J

2/25/03

Set	Items	Description
S1	975	DATA(W)MINING AND BUSINESS AND DATA(W)ANALYSIS AND (STOR? OR DATABASE) AND REPORT
S2	704159	(WEB OR NETWORK OR INTERNET) AND SEARCH? AND (MARKET? OR FINANCE OR SUPPORT OR RESEARCH(W)DEVELOPMENT OR SALES OR EXECUTIVE)
S3	30422	SEARCH(W)ENGINE AND (HIGH(W)TECHNOLOGY OR ELECTRONICS OR AUTOMOTIVE OR FINANCIAL(W)SERVICES OR ENTERTAINMENT)
S4	21	S1 AND S2 AND S3
S5	11	RD (unique items)
S6	9	S5 NOT PD>010207
S7	16382	MINDSHARE
S8	0	MARKET(W)METRIC(1W)ANALYSIS(W)SYSTEM
S9	233	BIZ360
S10	101	S7 AND S9
S11	40	RD (unique items)
S12	40	S11 NOT PD010207
S13	0	S11 NOT PD>010207
S14	0	AU=(TSANG, YOU? OR TSANG YOU?)
S15	0	AU=(GARTUNG, D? OR GARTUNG D?)

?

01730804 03-81794

Mine over matter

Baker, Sunny; Baker, Kim

Journal of Business Strategy v19n4 PP: 22-26 Jul/Aug 1998 CODEN: JBSTDK
ISSN: 0275-6668 JRNL CODE: JST
DOC TYPE: Journal article LANGUAGE: English LENGTH: 5 Pages
WORD COUNT: 3030

ABSTRACT: Just 5 years ago, only 50 researchers received Gregory Piatetsky-Shapiro's monthly electronic newsletter, Knowledge Discovery Nuggets. Today, the newsletter has 4,000 readers and is published 2 or 3 times a month. **Data mining** - the use of statistical methods and new **search** software to uncover useful patterns inside databases - continues to attract more attention in the **business** and scientific communities. There are massive amounts of data in corporate coffers that could be used to reinvent **marketing** strategies, and **data mining** is one way to find information that matters.

TEXT: Headnote:

There are obscene amounts of data In corporate coffers that could be used to reinvent **marketing** strategies. **Data mining** Is one way to find the information that counts. 1

PITY THE **MARKET** RESEARCHERS FROM PREVIOUS decades. Blessed with ungainly **database** design, slow throughput, and a host of compatibility problems, gathering even simple customer demographics required days-long model-building exercises to launch only the crudest **search** for useful information. Assembling a data sieve required help from the MIS gurus who, with a Dilbert and Wally level of **marketing** savvy, failed to understand the purpose of pulling facts from the data archives. Underfunding the **data mining** effort was common because what senior managers assumed would take minutes really took weeks to plan, implement, test, analyze, and summarize. 2

The resulting multi-page reports were of more use to paper recyclers than to information-hungry decision makers. In those dark days, **data mining** had much in common with coal mining. It was dark, awkward, gritty work with perilous hazards that could endanger the life of the **business** . 3

Just five years ago, only 50 researchers took part in a knowledge discovery and **data mining** conference workshop and also received Gregory Piatetsky-Shapiro's monthly electronic newsletter, Knowledge Discovery Nuggets. Today, however, the newsletter has 4,000 readers and is published two to three times per month. And **data mining** -the use of statistical methods and new **search** software to uncover useful patterns inside databases-continues to attract more and more attention in the **business** and scientific communities. 4

According to Piatetsky-Shapiro, still the editor of Knowledge Discovery Nuggets and also a director of applied research at Knowledge Stream Partners, a **data mining** consulting company based in Chicago, "We are clearly moving into the next generation of **data mining** systems, toward more and more embedded solutions." Over the last two to three years, there has been more emphasis by developers on integration, visualization, and data access tools, and the number and sophistication of these tools is increasingly rapidly. And in a 1997 **report** , Stamford, Conn.based Gartner Group stated: " **Data mining** and artificial intelligence are at the top of the five key technology areas that will clearly have a major impact across a wide range of industries within the next three to five years." 5

Companies now use computers to capture details of **business** transactions such as banking and credit card records, retail **sales** , manufacturing warranty, telecommunications, and myriad other transactions. The data from these transactional systems contain thumbprints of the key trends that 6

affect various aspects of each **business** -including information about products that sell together, sources of profits, factors that affect manufacturing quality, and other mission-critical relationships. The predictive relationships and trends buried in the data are the gold revealed through **data mining** .

Fueled by the increasingly rich data sets maintained and often underused by corporations, a new generation of **data mining** tools from companies like AbTech, **Data Mining** Technologies, Future Analytics, Information Discovery, NeoVista, Ultragem, and others, are moving **data analysis** out of the MIS department and closer to the boardroom. While it's true that a complicated **search** still demands the services of a data analyst or two, new tools enable managers in many companies to conduct much of the work by themselves with only a modicum of training to make wise and informed decisions. 7

Today, the newer and easier-to-master **data mining** tools can become a senior manager's Rosetta Stone for translating even hundreds of terabytes of data into useful models (a terabyte is a 1,000 gigabytes, and even one is a lot of information to sort through). A dig in the right place with the right tools translates flat **database** gibberish into gems of insight about the organization and its customers. 8

Tomorrow Is Today

Because large databases often provide too much of a good thing, approaches based on traditional query languages and OLAP (on line analytical processing) usually encounter a problem known as "The Maze of a Million Graphs" where a user can build a million charts and still not see the forest for the trees because there is so much data. **Data mining** , on the other hand, draws its power from software tools that **search** through the data intelligently to look for patterns and relationships. The best software, on its own initiative and using embedded algorithms, discovers key patterns in the data warehouse by itself. 9

1C **Data mining** tools extend decision **support** capabilities that allow managers to query information in databases and turn the results into reports. The difference is that **data mining** tools look at vast amounts of data, often from multiple sources, and find patterns and relationships in the data that would otherwise remain obscured from ordinary **database** queries. 10

1A 1D 1B **Data mining** goes hand-in-hand with the new data warehousing tools that are necessary to organize the historical information gathered from large-scale client/server-based applications. For example, **data mining** tools offer a way to identify, extract, and analyze the valuable information contained in the massive databases that are often a byproduct of on-line transaction processing (OLTP) systems. The data from the on-line transactions, after being appropriately "cleansed" and organized in a relational **database** format (commonly referred to as a data warehouse) can then be mined to discover patterns in customer activity. These patterns provide **business** analysts with insights into customer behaviors that would otherwise be indiscernible. The insights lead to proactive, optimized **business** initiatives to reach, serve, and target customers. 11

What Can You Find in a **Data Mining** Expedition?

1C **Storing** embarrassing amounts of data is something that computers excel at. Getting it back is a different matter. Without intelligent software intervention, snagging the right data is impossible. And only within the past few years have easy-to-master **search** tools become affordable and powerful enough to get information onto ordinary desktop computers. 12

The best **data mining** tools generally solve one or more of the four basic problems in turning warehoused data into predictive information: 13

Classification of data into meaningful entities.
Classification results in defining the data elements that exhibit similar behaviors or relationships to other data.

Modeling explicit dependencies between variables, such as the dependent relationships between income and nonessential purchases, to cite a simple example.

Clustering data into a finite set of categories (clusters of data variables) that describe the data relationships in terms of predictable actions, performance, or behaviors.

Detecting deviations in key data from previous or expected values, which allows data miners to use deviations from the past to predict changes in future trends, output, or behaviors.

Data mining tools that solve these data problems produce predictive knowledge from data warehouses. For example, in customer prospecting and segmentation applications in industries such as banking, credit cards, and insurance, **data mining** helps break the **market** into meaningful clusters and segments. This helps identify customer groups that are not only more likely to respond to offers, but also will provide better profits in terms of higher volumes of product usage. And, by analyzing results of direct mail campaigns, **data mining** helps identify higher-value responses (in terms of profit per mailing) as well as higher response rates. This information helps **marketing** managers and **business** executives focus their promotional activities and new **marketing** campaigns on the opportunities with the biggest payoff.

In customer relationship management, **data mining** finds patterns of product usage and consumer behavior. It helps managers understand what causes customer attrition and what will improve customer retention. It uncovers crossproduct utilization behavior, improving the management of channels such as bank branches, automated teller machines, and service outlets. These patterns can reshape the thinking of a customer loyalty department. And, because **data mining** can combine trend analysis with wallet-share and affinity results, companies can design better **marketing** programs that can be implemented based on a customer's life cycle model.

In the retail sector, **data mining** helps identify how products sell together in specific **stores** or regions through affinity and **market** basket analysis. **Data mining** also helps companies understand profit patterns per linear foot of shelf space, based on **store** layouts and product combination offerings, and determines when promotions work best and which item layout combinations are the most profitable. **Data mining** also helps identify products that are traffic builders and those that are profit makers. Retailers can assess the impact of advertising and in- **store** promotions and more accurately analyze inventory control issues.

Data mining efforts in companies specializing in consumer packaged goods are the mirror image of retail data analyses. These **data mining** expeditions analyze one manufacturer's products in many **store** chains, instead of several manufacturers' products in one chain. Here, **marketshare** analysis by **store** -chain and demographics, and understanding where and when promotions work and when they don't helps focus advertising dollars. And, by analyzing historical performance from the chain level down to the region and **store** levels, companies can identify the factors influencing the effectiveness of advertising.

Manufacturing quality programs can use **data mining** to help manage interactions among a large number of variables. **Data mining** tools automatically identify the unusual data densities that are the tell-tale signs of process variations within manufacturing and assembly operations. For example, companies can automatically identify the combination of plant, build-date, and product models that have a higher malfunction rate, allowing quality engineers to quickly focus on the source of problems. **Data mining** can also help improve product quality by identifying the factors that give rise to problems.

In the telecommunication industry, **data mining** identifies the patterns of change in the **market** , allowing the **marketing** department to better

focus on customers who demonstrate a high acceptance of services and longer usage. **Data mining** also helps with capacity planning by providing an understanding of the underlying patterns and structures of service usage by customer groups. This insight allows capacity planners to optimize the investments in **network** facilities to better serve customers, while avoiding costly overexpansions.

What Do You Need?

All the **data mining** applications mentioned already, and more, including programs for direct **marketing** , fraud detection, and stock **market** prediction, depend on reliable software programs that integrate with the corporate data standards. The program or suite of programs you choose should be fast, adaptable to your specific **business** needs, easy to learn, simple to use, and able to present results in terms that non-statisticians can understand. You should also buy the software from a company that has a good reputation for **support** , training, and consulting services, because even the best **data mining** software needs explanation. The vendor you choose should be willing to show you the ropes for going down into the data mines.

Beyond the right software, there are four additional components required for successful **data mining** :

1. Quality data organized into an accessible and extendible data warehouse. The data must be of the right age and richness (depth) for the task. Older data is often useful for studying trends, so age is not necessarily a criteria that renders it useless, even though the same data might be almost worthless when assembling a simple mailing list.

2. A well-executed model for extracting the data. While most mining operations aren't that difficult to assemble, a mistake is a two-edged sword. Snag the wrong data and the results may appear credible while completely missing the most valuable relationships among the records. Grab too much data in a model and the results become too cumbersome to produce a useful study.

3. A way to verify models and results. Testing preliminary results is a must for ensuring the project doesn't run off the rails for something a simple as a syntax error or as complicated as flawed logic. Analysis and testing of the completed results is also a must. A good **data mining** software company will provide access to training and consulting services to assure that your **searches** for gold in your data produce the pay dirt you're looking for.

4. An accessible reporting function. The output from **data mining** software should make a persuasive case for the results of the study. The best reports incorporate statistical data for analysts, coupled with a simple but convincing summary of data implications for less technical management.

Companies moving into **data mining** must first inventory the databases they own or have access to. Understanding what's already on tap spurs ideas on integrating the information for problem solving and customer management. Many companies have more data than they realize, often spread among departments as well as the MIS facilities. An inventory and merger of these data sources may result in eliminating previously duplicated databases and correcting a surprising number of mistakes.

There's no need to rely on internal data alone, however. Today, companies can buy, rent, or even borrow data from so many sources it's mind boggling. A major pizza chain looking to open new outlets might rent data from one of the many demographic data providers. A large corporation might combine its own data with commercially available data to analyze **business** customer needs and assemble a model to study historical trends and, from that study, decide which of several product prototypes to fund.

Industry surveys indicate that over 80% of Fortune 500 companies believe

that **data mining** will be a critical factor for **business** success by the year 2000. Most of these companies now collect and refine massive quantities of data in data warehouses.

These companies realize that to succeed in a fast-paced world, **business** users need to be able to get information on demand. And, they need to be pleasantly surprised by unexpected, but useful, information. There is never enough time to think of all the important questions. You need a computer with **data mining** software to help find answers to the questions you don't have time to ask. Such **data mining** can provide the winning edge in **business** by exploring the databases on "automatic" and bringing back invaluable nuggets of information. 27

If you're not already involved in **data mining** , it's time to get out your computer-based pick ax and start digging for the gems in your data. And if you don't, you can be that someone else will. 28
Sidebar:

Data Mining Software Companies

The following is a short list of companies that provide **data mining** software solutions For more companies and information on **data mining** , use the keyword **data mining** ! with the Yahoo! **search engine** on the WorldWide 29

AbTech Corporation 1575 State Farm Boulevard Charlottesville, Va. 22911-8411 limB: 804-977-0686 FaX: 804-977-9615 Email: **sales** @abtech.com 30

AbTech's ModelQuest sup (R) mining tools are based on a number of advanced machine learning technologies and AbTech's proprietary StatNet Expert approach. StatNet Expert technology is a **data mining** approach that has evolved from more than 30 years of research in developing predictive data modeling solutions to real-world problems. It combines the neural net, regression, and expert system methods and automatically captures the complex, non-linear relationships in data. 31

Data Mining Technologies, Inc. 1500 Hempstead Turnpike East Meadow, N.Y., 11554 Phone (516) 542-8900 Fax: (516) 794-4672
Email: info(at)data-mine.com 32

Data Mining Technologies is a new company that produces Nuggets, which automatically sifts through data and uncovers hidden facts and relationships. Nuggets can reveal which indicators most affect your **business** , and help predict future results. It works across industries and is effective with a wide range of applications. According to the company, it can help identify the best **market** segments, predict product success, reduce fraud, assist with credit scoring, forecast equipment failures and maintenance needs, improve manufacturing quality control/fault analysis, and assist with medical studies and scientific research. 33

Future Analytics, Inc. 7 West Washington Street P.O. Box 1455 Middleburg, Va. 20118-1445 Phone: (540) 687-3692 Fax: (540) 687-3654 34

Email: info@futureanalytics.com

Web Site: http://www. futureanalytics.com/ Future Analytics, Inc. provides cutting-edge analytical services to **business** users. Its core technologies consist of **data mining** , statistical analysis, data warehousing, and **Web** enablement. The company uses technologies such as neural networks, genetic algorithms, and decision trees to solve complex analytical problems. The company's professionals help companies design research studies, create data collection forms, and carry out statistical analyses. This service can be combined with **data mining** and data warehousing components. The company 35

Sidebar:

provides an integrated service, from warehouse assessment ants conStruction

to warehouse exploitation.

Information Discovery, Inc. 703-B Pier Avenue Suite 169 Hermosa Beach, Calif. 90254 Phone: (310) 937-3600 Fax: (310) 937-0967 **Web** .. http://www.datamining.com

Information Discovery, Inc is a leading provider of largescale **data mining** -oriented decision **support** software and solutions. Its products serve all the major decision **support** needs with pattern discovery and **data mining** software, strategic consulting, and warehouse architecture design. The company offers a variety of customized **business** solutions to augment the **Data Mining** Suite and the Knowledge Access Suite products for enterprise-wide, large-scale decision **support** . Using these products, companies can directly mine large multi-table SQL databases with no need for sampling or extracting files. And the Knowledge Access Suite includes a gateway to knowledge that has been pre-distilled from data and **stored** in a pattern-base. **Business** users need not perform **data analysis** , but simply query explainable knowledge on the intranet that has been automatically pre-mined.

NeoVista Software, Inc. 10710 North Tantau Avenue Cupertino, Calif. 95014 Phone: (408) 777-2929 Fax: (408) 777-2930 EIa:webmaster@neovista.com

According to company representatives, NeoVista Decision Series' knowledge discovery methodology can expose the patterns that are critical to **business** success to provide an advantage the competition may never discover. NeoVista's Decision Series suite of **data mining** tools enables companies to detect the patterns and trends in corporate data that lead directly to predictions of customer behavior, a targeted **marketing** focus, improved operational effectiveness, and optimal return on investment. NeoVista's Decision Series products integrate with open relational databases and standard SMP parallel hardware platforms.

Ultragem **Data Mining** 450 Wildberry Drive Boulder Creek, Calif. 95006 Phone: (408) 338 3302 Fax: (408) 338 7503 EmI:mail@ultragem.com
Ultragem uses the company's proprietary high-performance genetic **data mining** technology. Genetic **data mining** is the automatic extraction of prediction and classification rules from databases using advanced algorithms supported by the Ultragem software. For example, **data mining** can discover rules that will accurately predict the probability that a borrower will default on a loan. Genetic **data mining** can discover prediction rules in data which no human mind could ever have found, and which were beyond the reach of earlier technologies.

Author Affiliation:

Sunny Baker, Ph.D., is associate professor and director of technology for East Tennessee State University. Kim Baker, a **marketing** consultant for **high - technology** companies, is the author and co-author of more than 20 books.

GEOGRAPHIC NAMES: US

DESCRIPTORS: **Data mining** ; **Market** research; Statistical methods; Data bases
CLASSIFICATION CODES: 9190 (CN=United States); 5240 (CN=Software & systems) ; 7100 (CN=Market research)
?

6/9/4 (Item 2 from file: 16)
DIALOG(R)File 16:Gale Group PROMT(R)

06938168 Supplier Number: 58545238 (THIS IS THE FULLTEXT)
The Answer Machine.(information services management)(Industry Trend or Event)
Feldman, Susan
Searcher: The Magazine for Database Professionals, v8, n1, p58
Jan, 2000
ISSN: 1070-4795
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Professional
Word Count: 12581
TEXT:
When we **search** for information, we want answers, not documents. Current retrieval systems find documents that may or may not contaln the answers to the questions users ask. In the next 5 years, perhaps sooner, information systems as we know them will change dramatically. These systems will find real answers, moving from the static to the dynamic, using machine learning techniques to adapt to new information and to new interests. Finally, these systems will learn to interact with the user, delivering information in visual, easy-to-understand packages that can be manipulated and used collaboratively.

Datasearch

This information revolution is fueled by increased demand, by improvements in computer technology, and by our growing comprehension of how people seek and use information. As non-information professionals have become the dominant information consumers, they have begun to demand systems that can locate and manipulate information without arcane command languages and other traditional priestly rites. Unlike information intermediaries, whose main function is to **search** , knowledge workers use **searching** as a means to an end. This increasingly sophisticated group of information end users needs to find the right information quickly, analyze it, combine it into reports, summarize it for upper management, or use it to make decisions. They need a suite of integrated, intelligent information tools to make sense of today's ceaseless information bombardment.

Faster, bigger, cheaper desktop computers have the capacity to run newly developed information handling tools. News information systems will be built upon a foundation of linguistic analysis of language and meaning. To this, we add our growing understanding of cognitive processes. Research into how people think, combined with observations of how they interact with computer systems, is spawning the new discipline of human-computer interaction. New systems will draw heavily upon this field, as well as on cognitive psychology, graphic design, linguistics, computer science, and library science, each system with its own unique perspective on how to organize, find, and use information effectively.

The growth of corporate intranets adds to the demand. Companies are willing to invest in high-end, carefully crafted systems. **Business** cycles are growing shorter, while pressured employees spend too much time trying to handle too much information. Knowledge walks out the door as employees leave for new jobs in other companies. Intranets attempt to preserve this information and make it available to the entire company and will become an interactive venue for working with colleagues and with information in one smooth process.

Today's document retrieval systems lump all information needs into a single process. New information tools will separate these different needs into categories and provide specific tools for each kind of need.

Here are some of these **search** types:

* Broad subject **searches** -- fishing expeditions about a topic unfamiliar to the **searcher** . Appropriate terminology is hard to determine at first.

* Narrow, well-defined subject **searches** on a familiar topic with known terms.

* Comparative, information-seeking -- which company is the biggest, has revenues of more than $X, or more than 100 employees?

* Known-item **searching** for a specific title, author, or publication.

* Continuous monitoring of a subject.
* Pattern matching for emerging trends: foraging for matches to a description of an event or a profile of a competitor or other entity.
* Fact or statistic location,-- who, what, where, when, how?
* Chronological reconstruction of events or actions

The **Search** Process

How do we **search** for and use information? Do end users differ from information intermediaries, and if so, why? Can we differentiate types of **searches** and develop specialized tools to improve our finding and use of information? These questions and more must be answered as we Set about designing the next generation of information systems.

Searching isn't linear. We know that people engage in an iterative, or circular process when they seek information (see Figure 1 on page 61).

After testing the **search** behaviors of both end users and information professionals for the last 5 years, I believe in the inherent differences between how both groups **search** . This is not surprising, but it has little to do with the skill or training of either the information professional or the end user.' Rather, these groups differ in their fundamental motivation for **searching** . End users know why they are **searching** , even if they don't articulate their information needs well. Success is defined by an answer to their information needs. They will know it when they see' it. Therefore, they will more likely enter a very broad query, and then browse. In fact, given a choice, they will enter the **search** cycle by browsing first and then refining their browsing with a query. This explains the popularity of directory sites like Yahoo!

In contrast, the intermediary has only the end user's question to match. Success is defined by the best possible match. Therefore, intermediaries focus on precision. Their queries tend to be much narrower and they will **search** before they browse. A broad query to the information professional is unprofessional, sloppy. When we criticize end, users' for their lack of **searching** artistry, we are' often mistaken. They need to browse and browse broadly (see Figure 2 on page 67).

Most of today's document retrieval systems match queries to documents.

These systems address the middle of the information-seeking process, en closed in the dotted lines. While we may complain about the results, in fact, the systems do a pretty good fob of matching the actual query received. However, the systems ignore the two outer ends of the process, offering no help at all in translating information nets into questions and: then into acceptable queries. The systems do little to help the user understand and analyze what the system returns. So, while the user actually receives some good matches to his query, the query rarely reflects the information need behind it.

Yet, if the information need is not represented accurately, then the results returned will at best intersect that need spottily. Today's information sys terms require the **searcher** to extract terms that have the best chance of representing the question, while at the same time, eliminating extraneous or unrelated documents. We usually resolve this dilemma by using lists of nouns or phrases that represent the concepts in the question. In the process of formulating a query, we eliminate the actual meaning of the question because we strip away the context.

Look at the list of questions in the "Stinkers" sidebar at left. A real Answer Machine could answer these questions, and more. It should:

* Help the user formulate a query.
* Find answers, not just documents.
* Anticipate user needs.
* Retrieve data in any format, from multiple sources and suppliers. and merge the data into a single, de-duped retrieved set.
* Provide analysis and reporting tools to manipulate retrieved data.
* Display answers in easy-to-digest visual formats.
* Find patterns within data to **support** decision-making.

This is not as far-fetched as it sounds. Most of the, technologies that the Answer Machine requires are already in development. Answer machines will become the technical underpinnings of knowledge management systems, providing single, organized, easy access to all the information in an. organization including the following:

* Internal documents and databases in a variety of formats
* Committee reports

* External sources
* Directories of people and skills
* Tools to manipulate information and extract new knowledge

The trick will be to select the appropriate tools and then to present them as a seamless system. All the technologies discussed in this article should be thought of as pieces of a whole: a new model for a information system that brings together all of the resources of an organization in any format. 14

If you approach your information system as a whole, then you will implement each new technology as a brick within an entire edifice. You could implement each technology separately, but ultimately, integration, of these technologies will create a knowledge management system arid even a decision **support** system. Without this vision, you may end up with so many oddly sized bricks that you will have to start again from scratch. 15

The system you build should adapt to user needs and integrate information in any format. It must reveal patterns and trends in information, because patterns, and trends are usually more significant than discrete facts or nuggets. And above all, it must deliver answers to questions. 16

The Foundation

Any retrieval system must distinguish between one document and another. The system relies on indicators that determine what a document is about. It also tries to differentiate between documents "mostly about" a topic and those merely "somewhat about" a topic. Unique terms or phrases often serve as good discriminat6rs. However, unique terms are hard to find in some areas, such as **business** , which use very common words to mean something quite precise. The sample queries "Stinkers" offer good examples of this problem. 17

To best determine a document's meaning, ask a subject expert. Indexers do this f6r a living: However, while experts may agree On broad subject areas, they may differ On which terms to assign to a specific document. So the studies done on indexer consistency found that indexers assign the shine term to the same document only 50 percent of the time. Indexers do classify documents in the correct general subject area, even if they don't assign precisely the same term. They don't put financial institutions under environmental science. 18

Why, then, don't we stick to human classifiers to determine what a document is about? There are several reasons. First, that 50 percent consistency rate is quite troubling if **searchers** use thesauri to aid in query formulation. 19

Assigning the wrong term can eliminate a highly relevant document from a retrieved set. Second, human indexing is slow; it adds weeks, even months, to the time it takes to make something available online With real-time publishing becoming an accepted practice, we need other reliable means of distinguishing the relevant from the irrelevant. Third, the sheer volume of information is too great to try to classify it all manually. 20

Given that we must find an automatic means to select the best documents for a query, how can we teach a computer to recognize a good match? 21

Statistics and Probability

For all that **searchers** talk about words, terms, commands, and other linguistic phenomena, computers really understand only numbers. Every ASCII character, every letter in the alphabet, must be translated into a sequence of ones and zeroes before a computer can crunch it. Boolean commands work quickly because they are mathematically based. One of the ironies of online **searching** is that, its practitioners consider themselves to be "word" rather than "math" people. Yet, they handle Boolean logic with aplomb. 22

The genius of people like. Gerard Salton lay in their recognition that text contains predictable patterns. These patterns can be described mathematically, so that computers can detect them and then perform statistical and mathematical operations on them. For instance, it seems, obvious that the more a document is "bout" a subject, the more times words dealing with that subject will appear in the text. Conversely, these terms should not appear very frequently in documents not "about" that subject. This is the rudimentary idea behind relevance ranking in retrieval, systems. 23

Clusters of certain terms are even better indicators, that a document is about a particular subject. The appearance of co-occurring terms will 24

determine more precisely when a topic is central, to a document. None of this requires that we understand the meaning of the words, merely the patterns the words display in the text.

Needless to say, we could embellish this principle by saying that words in the title are more important than words in the body of the document. We could add that the closer together subject-relevant words appear, the more likely the document is about what we are **searching** for. Or, if the words appear in a lead paragraph, they are more important indicators of the subject than if they appear in paragraph five. This is what skilled **searchers** do in crafting a **search** . It is not magic.

If we can describe these patterns, we can program a computer to find them. The first mathematical operation that **search** engines do is to count, something that computers do very well and very fast. Computers count the number of times a term or terms appear in a document, then assign a weight, or number, that represents this count to distinguish one document from another. This weight calculation, usually takes into account how rare the term is in the whole **database** -- how many times it appears in every document in the collection. Rare terms are often good discriminators and receive a higher weight.

Search engines may also truncate terms to include plural and singular forms. Extra weight often attaches to terms appearing in the title or lead paragraph, as to documents which contain several query terms in the same sentence or in the same paragraph. Most **search** engines also, "normalize" results to take into account variations in the length of documents, since longer documents will probably contain more occurrences of a term. When a **search** system matches your query terms to documents, it adds up the weights for each query term that appears in a document and assigns a score for that document. Then it compares all, the scores and presents the highest first. This is relevance ranking in a nutshell.

Statistics and patterns enter into advanced retrieval systems in a number of other contexts. For instance, in order to determine whether a document matches a query, the system must calculate the similarity of the document to the query. The human mind does this without trotting out an algorithm. Computers must translate both query and document into some sort of representation. About this task, experts have written whole books.

One approach is to translate both query and document into a "vector" - a line which goes off at a specific angle from the center of an imaginary space. Think of this space as having a signpost at the center, with each individual sign pointing in a slightly different direction. The words in the document all point to specific directions in this imaginary landscape. Documents containing similar words will point in the same general direction; the more similar those document terms, the closer their angles will be to each other. We can measure these angles to give us a degree of similarity. This "vector space model" can help calculate relevance ranking, but it can also determine clusters or clumps of similar documents. This is the basis for most of the star maps or imaginary. Landscape visualizations used to display the contents of a **database** , or a retrieved set of documents.

These, statistical techniques work surprisingly well in the majority of cases. But these techniques do not work well for every query. That is the nature of statistical methods. When we hit an exception to the rule, the errors can be, glaring, unlike human errors. For instance, when a query contains both a very important concept expressed in an extremely common term and a very minor concept expressed in a rare term, then the rare term may skew the relevance ranking, since it has a higher, weight than the common term.

Remember also that statistical systems do not "understand" a query, but operate on the numbers. Many meanings for the same word elude this kind of technology Financial institutions may be classified as environmental science, if the word is "bank." However, since bank will not appear in combination with other environmental terms, if a query is more than one word long, a statistical system Would rank such a false drop low. Hence, **search** engines look very stupid by making errors that any human with half a brain would never make. This could explain why **search** engines have such a bad reputation among most professional **searchers** ; their errors are unreasonable. That is because the meaning of the terms being retrieved is not part of the equation for statistical processing.

Natural-Language Processing

In order to build a state-of-the-art information system, one must extract as much meaning as possible from each document. A list of words, or even words and phrases, is not enough. Context and meaning must be preserved. Only a system able to distinguish meaning can return articles about terrorists instead of rugby matches when asked for attacks, skirmishes, and battles in Rwanda. A meaning-based system will also know to return predictions about future, not past, production of widgets in Zambia in Question 9 of our "Stinkers" list.

To create an advanced information system, first one must build a knowledge base. This base will contain all the documents in the system and their words, but also added information to resolve meaning and dissolve ambiguities. A good natural-language-based system provides the foundation for this system, because it parses sentences thoroughly, extracts meaning from context, and is smart enough to realize that if the year is 1999, Hilary Rodham Clinton and the first lady are the same person. A document-processing tool is required that can extract and **store** many layers of meaning, as well as automatically categorizing documents and identifying all variants of proper names. Each unit of meaning may also carry a time stamp relating to the content, not to the date on which someone added the document to the **database** . With relevant dates in place, later tools can extract automatically chronologies of events. Chronological information also enables the system to distinguish between first ladies Barbara Bush and Hilary C linton, depending on the time and context of the question. The information in the knowledge base should also be retrievable as separate units, such as a single sentence or paragraph, if we want it to supply direct answers to questions.

I stress this knowledge base building step because most organizations will not willingly invest the money, time, and effort needed to design a knowledge base more than once within a few years. Any future advanced information tool will operate on the contents of this knowledge base. Therefore, extracting as much knowledge as possible should increase the flexibility in the future to adopt new technologies, as they arrive We can't know now) what tools in what formats; research and the **market** will deliver in the next 5-10 years Compatibility will always be an issue However, raw knowledge does not change. The more handles that you create to grab a Piece of information, the more chance that you can retrieve it when needed. This is the same principle that advises digitizing at a high resolution when scanning collections: Build the foundation wisely and richly, because you'll never be able to start again from scratch.

As we build advanced information systems, we will require that the systems understand text as we do. Natural-language (NLP)-processed-based systems are the only ones to answer this description at present. While NLP systems match terms, Las both Boolean and statistical **search** engines do, the systems also extract meaning from syntax, built-in lexicons, context, and even the structure of the text itself. This is what humans do to figure out what a document means.

Many people feel that statistical and NLP systems won't work as well on bibliographic databases because their forte is full-text **searching** . True, these systems are not designed to work well on document records that do not contain substantial text. Therefore, it is said that bibliographic records such as those appearing in typical library are not good candidates for these advanced retrieval systems. However, I have found these systems as effective as Boolean systems in **searching** through bibliographic databases, because most can default to a relaxed Boolean query if necessary As an added benefit the ability of the systems to relax the strictures of a query means that occasional typographic errors will be ignored in relevant records that a Boolean system would eliminate from the results.

Intelligent Agents

Imagine an information system that learned what you sought and began to anticipate what you would like to see. While this may sound like star Wars, in fact, this capability exists in embryonic form today. Interactions with today's systems are fixed in time. The **searcher** must change a query in order to find documents not already retrieved and to add new indexing terms manually. We need systems that adapt to both the changing interests of the user and to changes in the terms used to describe each topic. Machine learning techniques can make an information system dynamic.

For instance, suppose 3 years ago you set up an alert for anything on "information retrieval." If you didn't; change your Alert profile, you

would miss all the articles on **data mining** , knowledge management, or automatic summarization. An intelligent agent system could detect the rise of these new terms. The system would find clues in the appearance of **data mining** as a co-occurring term with information retrieval. Or, the agent system might note that you were reading articles on **data mining** and ask if you wanted to add that term to your profile. It might be programmed to follow new **Internet** links from sites that interested you; or it could run an updated query periodically on all the **Web search** engines and then follow those links. This is of immense importance in a world in which, in 1997, a Reuters survey found that most professionals spent more time seeking information than using it.

Intelligent agents are software programs that use machine learning. Agents do not have innate intelligence. Although agents can operate in situations that have underlying patterns or rules of some sort, agents cannot work in complete chaos or with random input. The patterns or rules that they rely on may be described by humans or developed by the agent-based system itself. An agent system develops rules from sets of representative data and queries -- a training set. During the training period, system agents "learn" the best matches by trying out various matches and receiving corrections from human input. Eventually, agents build a pattern for what constitutes a "good match." 39

Agent systems are autonomous -- in other words, they can initiate actions within a carefully defined set of rules. They are also adaptable, able to communicate with other agents. and with the user. Agents may be mobile, traveling along the **Internet** or other networks in order to carry out various tasks, such as finding or delivering information, ordering books, or monitoring events. Most importantly, agents can alter their behavior to fit a new situation. They learn and change. 40

Some agent systems exist today. See the Botspot (http://www.botspot.com) for an extensive list and description of such systems. The agents in the Microsoft Office suite are only a beginning. They are not adaptable and they follow set rules. These agents offer hints, take and sometimes answer questions about functions of the software, and are mildly amusing. Eventually, we can expect agent systems to adapt to our preferences for formats or other repetitive actions we take -- like opening applications in certain orders or checking e-mail at a certain time of day -- and will perform these tasks automatically. 41

Eventually, agents will play a big part in the decision **support** systems now in development. These systems will use a knowledge base to find and compare previous situations that might apply to current problems, offering alternative solutions and perhaps creating scenarios for each alternative. 42

These three disciplines -- statistics, natural language understanding, and intelligent agents -- form the foundation for understanding and using the information tools of the future. While it will be possible to use these tools and never understand their inner workings, those who delve below the surface rules will use them most effectively. Apparent anomalies and mistakes will be come less puzzling as well. 43

NLP-Based Technologies

By examining meaning instead of just matching strings of words, NLP

Systems can solve many retrieval problems intelligently. These include identifying concepts, even if different terms are used to describe the same idea. NLP systems should identify the names of people, places, or things in any form. The systems could also encompass speech processing, summarizing documents, and even groups of documents, and automatically indexing and classifying documents. Each of these aspects represents a distinct area of research with tools in development or, in some cases, already on the **market** . 44

Concept Extraction and Mapping

Concept mapping is the key to many new technologies on the horizon. Language provides rich alternatives in how an idea is expressed. Not only are there direct synonyms, but metaphors, similes, and other literary devices. These devices delight the reader, but puzzle the computer. We need systems that can use all those levels of language to interpret meaning correctly and to relate similar expressions of an idea to the same concept. 45

Concept mapping enables us to:

* **Search** across disciplines using different vocabularies to express the same idea. 46

* **Search** across languages for the same subject.

* Identify and retrieve all variants of a name or place, no matter how a question is phrased.

* Index materials automatically.

Concept and vocabulary mapping are like creating a controlled vocabulary. In a controlled vocabulary, all synonyms are identified and one is chosen as the "official" term. Other terms cross reference to that official term. Concept mapping works in a similar manner, except that the concept does not need to be a single chosen term. Instead, all synonyms form a cluster of terms that represent the idea. Since the idea is represented abstractly, it can cover not only words in one language, but in any other language and well beyond the conceptual grasp of multilingual human dictionaries or thesauri.

Vocabulary mapping, a form of this technique, enables a **searcher** using MESH terms in MEDLINE to **search** intelligently in CINAHL, another medical **database** with a different thesaurus in control. Thus, the idea of "tree" has multiple terms mapped to it, as shown below in Figure 3.

This is a technology already in place with varying degrees of sophistication. It is used in the following areas.

Machine-Aided and Automatic Indexing

Machine-aided or automatic indexing (MAI) finds major concepts in texts, maps them to an internal thesaurus or controlled vocabulary, and applies indexing terms automatically. It may also extract important names, disambiguate words, and identify new terminology for indexers to add to the system. MAI offers candidate terms to indexers for their approval. Automatic indexing applies these terms with no human intervention.

Machine-aided indexing has been around a long time. Most such systems are rule-based and assign terms based on rules such as "use 'automobile' as an indexing term whenever a document is about 'car'" just as professional human indexers do. Data Harmony/Access Innovations is well known for its rule-based machine-aided indexing systems. Northern Light uses rules developed by human indexers to automatically assign broad terms to all documents for its customer folders. Autonomy uses machine learning to automatically categorize materials, and Semio creates taxonomies or hierarchies automatically. Systems such as DR-LINK, developed by Dr. Elizabeth Liddy at Syracuse University, assign subject codes in order to disambiguate words. Some MAI systems work with up to 80 percent accuracy, which compares favorably with manual indexing.

Some experimental approaches use probability and statistics to categorize materials. Muscat, now owned by Dialog, is a good example of this approach. Others are experimenting with neural networks for automatic classification.

MAI systems can also extract important names from the text or "disambiguate" terms. Consider the term "bank." It may be a place to **store** money, the side of a river, a turn made by an airplane, or the slope of a curve on a highway or railroad. Increasingly **Web** and other **search** engines use automatic indexing to disambiguate or to create broad categories for browsing.

MAI can speed up the indexing and abstracting process needed to prepare databases. It particularly helps in handling such high volume tasks as assigning metadata terms to **Web** documents.

Automatic Summarization

Not too long ago, no one could find information. Now there is too much of it. Any tool that gets us quickly to the most important bits is valuable. Quick, automatically produced summaries have this potential. There are two kinds of automatic summarization. The first summarizes whole documents, either by extracting important sentences or by rephrasing and shortening the original text. Most summarization tools currently under development extract key passages or topic sentences, rather than rephrasing the document. Rephrasing is a much more difficult task.

The second process summarizes across multiple documents. Cross-document summarization is harder, but potentially more valuable. It will increase the value of alerting services by condensing retrieved information into smaller, more manageable reports. Cross-document summarization will allow us to deliver very brief overviews of new developments to busy clients. We can expect some tools to do this within the next 2-4 years.

Cross-Language Retrieval

Research communities now span the globe. Researchers need to know what goes on in their fields no matter what the language of the source, e.g., companies going global in scope and interest. Two approaches are in development. The first translates text from one language to another. The second maps words in the same language to a single coded concept, just as concept mapping does. Even rough wording or poor translation is adequate for cross-language retrieval. We can also use it for retrieving foreign language documents, even if we can't translate the documents perfectly. The combination of concept mapping and automatic summarization can deliver a rough gloss or overview of an article so that a researcher can decide whether to read an entire document.

Entity Extraction

Entities are names of people, places or things. As we all know, entities are often difficult to locate within a collection of documents because many variant terms may refer to the same person. For instance, "AT&T" may also be found as "AT and T," or "AT&T." "Marcia Bates" may appear as "Bates, M" or "Bates, Marcia," but should not be confused with "Mary Ellen Bates." President Clinton was once Governor Clinton was once Governor Clinton and still is Bill Clinton and William Jefferson Clinton, not to mention "the President."

Newer information systems develop lists of name variants so that all the forms of a name map to the same concept and will retrieve all the records, no matter which term appears in a query. These systems may also contain built-in lexicons with specialized terms and geographic name expansions, e.g., to include France when the **searcher** asks for Europe. System administrators should have access to the lexicons to add internal thesauri and vocabulary. They should also add new names or terms as they occur in new materials. NetOwl is one example of a product that extracts entities. For decades, LEXIS-NEXIS has used name variants in order to improve retrieval, but automated extraction and **storage** give this policy far more power.

Relationship Extraction

With extracted entities in hand, one can perform some interesting analyses across documents. For instance, one could find out who has met with whom over the timer period of the collection. This kind of **data analysis** requires that the system extract relationships among entities. Some systems can extract more than 60 different types of relationships, including some that describe time or tense and numbers. Natural language researchers have developed categories to describe these relationships. For instance.

* The ISA relationship defines who or what the subject is: "Gil Shahan is a fine violinist."

* The AGENTOF relationship describes who or what caused an event to happen or had causal relationship: "Increased ozone in the Southern Hemisphere causes severe sunburns."

Tools like KNOW-IT, developed by Woojin Paik of Solutions United, extract entities and **store** their relationships to each other. This involves a larger chunk of information than single words or even phrases, consisting of the subject, the object, and the kind of relationship they have to each other. That way, we know who initiates what action and what its effect is on whom. These tools would **store** Jim owes Fred as a different unit than Fred owes Jim. The system can create webs of relationships that might help to direct which bacteria were becoming drug resistant as a result of which antibiotics or to detect which drug traffickers work together.

As we have seen, words by themselves often do not suffice to establish meaning. If one can **store** the context, the syntax, and the unambiguous meaning of each sentence as a unit, one can build a good question-answering system. Tools like this can answer questions such as, "Who fired the president of Consolidated Widget Company?"

Chronological and Numeric Extractions

If a system can determine when and what event Was happened, or howlarge something is compared to something else, then it can answer questions such as, "When was Netscape bought by AOL?" or, "Find all the Widget companies that produce -more than 5 million widgets a year." With this kind of information extracted from its contents, the system can also construct chronologies of events. This may not seem earth shaking, since one might find a biography of a person instead of constructing one, but

imagine the possibilities if the system could reconstruct the development of a competitor and then use that model to monitor news for emerging competitors before you have identified them.

Text Mining

Text-mining technologies differ from **searching** because they find facts and patterns within a **database** . In other words, text mining, looks at the whole **database** , not just a single document, and then extracts information from all the pertinent documents in order to reveal patterns over time or within a subject. These technologies perform some analysis on text in a **database** to present patterns, chronologies, or relationships to the user.

Librarians do **data mining** almost implicitly -- to them, information falls into patterns, groups, clusters, and hierarchies. While it may seem second nature to us, in fact, it is a rare talent. How can software accomplish the same thing? Well, it can't with any intelligence. But remember that language is made up of patterns; this fact lets us generate new but still under-standable, sentences. If you identify the clues that tell you, for instance, that something is a prediction, then the software can follow those same rules to find predictions, e.g., using terms like "by next year," "in 2010." Good text mining depends on the quality of the knowledge base on which it operates. If relationships, concepts, chronological information, and entities have already been extracted, then, the text-mining process can take advantage of this information and seek patterns within it.

Question-Answering Systems

We often lose sight of the purpose of information retrieval, which is usually to answer questions, not just retrieve documents. Question-answering systems look within documents or knowledge bases to find answers. For example, if you ask a question-answering system, "When was the Wye River Accord signed?," you will get an answer of October 1998, rather than a list of documents about the Wye River Accord, which may or may not contain the answer. Question-answering systems find the best matching answers extracted from within matching documents. If users need more information, they can link to the source documents.

Filtering, Monitoring, or Alerting

The difference between filtering and ad-hoc **searching** is that in **searching** , the **search** may change, but the **database** remains the same, while in filtering, the, **search** stays the same, but the data against which the **search** matches changes. Filtering only looks for new documents of interest. To set up a filter, the user creates, a profile or "standing query," which runs against any new additions to the **database** . The art of designing a standing query lies in creating a broad enough query to prevent the omission of important developments, while making it narrow enough to prevent too much information from flooding the user.

Like any other **search** technology, filtering or alerting depends on the quality of the **search engine** used. A **search engine** that can provide well-focused retrieval, preferably using some sort of disambiguation and concept extraction, will most likely catch related topics.

One of the major problems with any kind of standing, continuing query, or monitoring service is that the terminology in any field changes over time. So do a user's interests Yet, most of today's alerting services are static. Those who rely on profiles must make sure to update them regularly. As an example, my own 3-year-old alert on "information retrieval" returns very little of interest these days. Instead, I need to add **search** engines, **data mining** , text mining, filtering and routing, natural language processing, knowledge management, and many other new terms. Newer systems that incorporate some kind of machine learning or intelligent agents are vital for good continuing monitoring of topics. Filtering tools that incorporate machine learning can detect new terms and offer to add them to a standing query. They can also, note changes in the user's interests and adapt the query to fit these new topics.

Change Monitoring

Change monitoring is a specialized type of filtering. It monitors established documents or **Web** sites and determines when changes have occurred within them. The technique has become a vital part of competitive intelligence or events monitoring. If a competitor' **Web** site remains unchanged, the system ignores it, but it raises a red flag if substantial

changes and additions occur. Similarly, official agencies charged with collecting and archiving government documents need to know when a new revision of a form or document or law appears.

One company that monitors **Web** pages for changes is Ingenius Technologies (http://www.ingetech.com). Their JavElink monitors a list of URLs supplied by the client and reports only the changes. The visual display makes it easy. to note what has changed at a glance (see Figure 4 below). Ingenius also uses this technology to create emailed alerts (NetBrief, http://www.netbrief.com) that contain only the changed text of a site; The Ingenius site displays several free alerts on popular topics as examples. 71

A new extension of NetBrief sends a daily e-mail containing URLs and brief excerpts matching client keywords. Each day, InGenius reviews 100 online daily newspapers, as well as dozens of **business** and technical publications. Clients may add new sites or **search** engines as they wish. They may also specifically include or exclude certain sources or topics. 72

Visualization

The human eye understands visual representations, much faster than it can read-text. As? the old proverb says, "One picture is worth a thousand words." Compare the simplicity and speed of recognizing a picture of people sitting under a tree at a picnic to reading a description of the same. scene. In order to help people interpret large sets of data or documents, many researchers are designing visual equivalents of the text, so users can digest the information at a glance. 73

Visualization helps handle information overload. Imagine being able to Hand a one-page visual overview of the week's developments to the CEO of a company instead of a five-page digest. Visual information systems are also vital to crisis management, air. traffic control, and other situations in which people must respond instantly to a great deal of information. 74

Effective Visual representations are confined by the limitations of-the computer screen. There is, only so much information that can be displayed effectively on the standard 14-or 15-inch monitor. For an example of a nice. kind of interface to have see a description of the interface to Phrasier (http: //www.cs.waikato.ac.nz/-stevej/Research/Phrasier), an innovative system for browsing. by phrases. The screen design for this product is too large to fit a standard screen, but it contains all the elements that a user would want to have in order to interact well with an information system. It displays documents, related concepts, and key Phrases, all in one place. Figure 5 below shows part of the screen. 75

Most of the visual presentations of information we see today are experimental. We really don't know how people will interact with them. Cognitive psychologists, online experts, and computer scientists need more than the anecdotal information we get front us-ability tests in order to establish guide-lines for, good design. We do know that people have many different cognitive styles and that to interact with computers efficiently they need tools and interfaces that fit how they think. The great challenge will be to discover how the mind works and then to design tools based on this knowledge. 76

Some concepts are fairly simple to visualize effectively. Bar charts or even differently sized squares can illustrate quickly comparative sizes, amounts, or numbers. Timelines can show time-dependent events. Proximity of objects can indicate close relationships. Pie charts show how the parts make up a whole. When we move from these common concepts to representing relationship among people and places over time, then we must invent new imaging. 77

A visualization sits on top of the information retrieved from a system. While the interface determines how the information displays, what it displays depends on the data extracted. Thus, relevance rankings easily display as bar charts. The amount of information available on a topic can show as a set of colored boxes of various sizes. 78

The vector space model that we discussed earlier lies under most visualizations of subject content. It can create star charts, showing clusters of documents, or the imaginary land-form maps from Cartia). Look at this visualization of a set of **search** results from Cartia in Figure 6 below. The highest peaks represent subjects having the most documents. The closeness of hills shows proximity. 79

The browser from the Human Computer Interaction Laboratory (HCIL, http://www.cs.umd.edu/hcil/ndl/ndldemo/draft11/daveloc4.html) at the 80

University of Maryland gives an instant overview of the Library of Congress collections. As you pass your mouse over each timeline, it turns blue, and so do the types of collections that contain information about that time period.

Query formulation is one of the weakest spots in the information process. Several companies and research groups have developed visual aids to query formulation, but I still like the text power **search** screen from DR-LINK, developed by Dr. Liddy at Syracuse University, that shows you how the computer has interpreted your **search** and gives you a chance to change it (see Figure 7 below).

Spotfire (see Figure 8 below) and Dotfire, its newest form, are dynamic query tools. These tools present a set of categories that help to narrow down a **search** . You can manipulate each category using a slider. HCIL at the University of Maryland developed both of them (http://www.cs.umd.edu/hcil). Dotfire (http://www.cs.umd.edu/hcil/west-legal/dotfire.gif) is the new West-law case law explorer. (For more information, read the technical paper by Ben Shneiderman, David Feldman, and Anne Rose, "Visualizing Digital Library **Search** Results with Categorical and Hierarchical Axes," CS-TR-3992, UMIACS-TR-99-12, February 1999, ftp://ftp.cs.umd.edu/pub/hcil/03html/99-03.html.)

Figure 9 above shows the Hyperbolic browser from Xerox PARC, developed to help people explore the contents of a **database** visually You can find it at the InXight **database** . (http//www.inxight.com).

Gary Marchionini and his students at the at the Interaction Design Lab at the University of North Carolina study the effectiveness of interface designs for various kinds of resource formats, such as statistics or video files. The interactive statistical relation browser is a pro-totype developed for the Bureau of Labor Statistics. It displays, in one screen, subjects covered by the **database** , the number and format types for reports, as well as regions and dates covered. Related **Web** -sites also display. It is simple, effective. (See http://ils.unc.edu/idl/ for other research by this group.)

The Perspecta interface shows the user, in one screen, which parameters they can **search** . This screen shot also shows the results of a **search** done on their travel information **database** Each box shows the user, at a glance, the number of tours that exist in each of the categories requested during the time period indicated For instance, 87 canoeing tours are offered during a specific time By grouping results into logical bundles this software enables the user to understand the results of a **search** before be the to plow through actual hits:

Having tools that can give you several views of the same data helps you discover patterns.

Northern Lights custom folders give you a quick visual overview of **search** results The careful categorization of contents makes **searching** Northern Light both broad and well focused. Northern Light also **searches** Yahoo! directory pages. Yahoo! has some excellent resources, but I prefer to **search** rather than to start with a browse. Northern Light gives me the best of both approaches.

I like the simple display from TASC, (www. tir.tasc.com/Visualization/). TextOre shows the extent of the information about a subject by the size of the colored squares. If you click on a square, you will see the documents that it represents, or, for large document sets, further charts. This is visual **data mining** .

Tools to Analyze and Interact with Data

Finding and using information should be an active process. We need to read what we find, but we also need to merge sources, pull them apart, separate the data, into categories, sort the data, seek patterns, and send the information to colleagues and clients.

Puffin **Search** (http://www.puffin-ware.com) invites this kind of interaction. It **searches** across up to eight **Web** **search** engines at a time and brings the results back to your desktop. It saves the **search** results, creating a list of all the terms that appear in two or more citations. Then you can sort, cluster, and resort the results using any cell in the table as a basis for comparison. Choose a title and it will re-rank all other hits by their similarity to that title. Or, choose several of the keywords and rank all 1,200 hits by the terms you have chosen. You can sort by **search** **engine** or by URL. Puffin automatically

forms clusters based on the similarity of a group of documents, using a similar technique to the vector space model. You when you use it as a filtering tool.

Netbook, developed by the Human Computer Interaction Group at Cornell University (http://www.hci.cornell.edu),is part of a multimedia tools suite that foreshadows what the digital library will look like in the future. These are the tools that users will demand as we move to, dynamic use of information (http://www.hci.cornell.edu/projects/projs/multimedia.htm): 92

Netbook allows users to capture images from digital collections, **store** , and view them on a user's personal netbook page. Thumbnails of images can be organized, allowing users to build a manageable collection. 93

Annotator allows users to annotate or read annotations to images that have been taken from an online collection or a Netbook. A class of students for example can view the annotations an instructor has made to an image or create their own annotations visible to their classmates. 94

Authoring helps users manipulate, organize, and display research and data with hierarchical and hypermedia links. Students can view the work of others and organize andmake their own links. 95

Artview transforms online image collections from museums and other sites into collaborative learning spaces. At the same time but from different locations, users can view an image and communicate with each other using a shared text window. 96

Searching Multiple Sources Simultaneously

Searching across different kinds of information collections poses one of the biggest challenges facing digital library and intranet builders. Collections may encompass text or images or statistics. Text files may contain bibliographic records, abstracts, or full text. Image collections may only offer **search** engines the text appearing as captions. Once we move outside of controlled, integrated collections of the same kind of Materials, we encounter several obstacles. These include vocabulary differences, differences in type of materials, and differences in relevance-ranking algorithms. 97

Differences in vocabulary are a familiar problem to .any experienced **searcher** ; Each collection or source may use different terms to express the same idea. We professional **searchers** traditionally handle this problem by using every synonym we can think of. Thus, we might choose both pumps and impellers, or theater and theatre to round out a good query. In NLP systems, concept matching may perform some of this work for us. However, customized intranets may want to develop internal lexicons that would map pumps and impellers to the same concept automatically This is a good application for concept monitoring and indexing. 98

Searching across heterogeneous materials presents a Knottier problem, as **searchers** working with Dialog OneSearches can tell you. For instance, the weight of each word in a bibliographic record is probably enormously high compared to the same term appearing in a full-text, 10-page 'document. One could imagine trying to tweak a **search** system each time it adds a new kind of collection. 99

Searching across several systems complicates matters still further. Most **search** engines calculate the relevancy of a document by counting the number of occurrences of each query term in each document. The more occurrences, the more relevant the document. This works fine when the documents are approximately equivalent in length and of the same type. When we combine these materials in a single **search** , the results will skew by length of text. 100

If we try to **search** across **search** systems, as **Web** metasearch engines do, we find that each one measures relevance differently. In addition, since each system computes the relevance of a document to a query in part by finding out how rarely that term occurs in the **database** as a whole, and collection contains different materials, it is unlikely that what is highly relevant in one collection will rank the same way in another. Data fusion is a set of techniques for establishing a common ground to measure relevance. The lack of data fusion treatment explains why **Searching** across files in Dialog or metasearching on the **Web** doesn't work well within relevance-ranking systems. 101

Here's an example. Suppose that: we decide to **search** for a few good articles on the causes of high blood pressure. We pick two **Web** **search** engines. But, we don't know that **Search** **Engine** 1 covers all the major medical information sites, while **Search** **Engine** 2 concentrates on 102

sports. **Search Engine** 1 finds 250,000 articles about high blood pressure. It ranks them. **Search Engine** 2 finds 10 articles, and they have only minimal information on the subject. Think back to our weighting algorithm. If high blood pressure appears rarely in a **database** , it gets a high weight. SO, **Search Engine** 2 gives all of these documents a 98 percent ranking. Since high blood pressure constitutes a common term in **Search Engine** 1, it gets a lower weight. If our metasearch engine takes the top 10 from each, we will see all 10 of the **Search Engine** 2 documents before we ever get to those from **Search Engine** 1. Yet, the results from **Search Engine** 1, coming from medical sources, may be vastly superior.

Data fusion tries to merge results from several **search** systems. One technique takes one document from each in a round-robin approach. Another creates a virtual collection that merges all the documents found in all the databases. Then weights are reassigned based on this common collection. The second technique gives better results, but is computationally more costly.

103

Evidence Combination

Evidence combination improves retrieval from the same collection by using different retrieval techniques. It will be a hot topic in the next few years, as computing power increases still further. Any retrieval technique is faulty and will omit some relevant documents, perhaps due to a poor query, to differences in terminology, or even to errors in spelling introduced by optical character recognition programs. **Searchers** may also miss important documents if the documents do not appear in the top 30 or 50 examined. Certain ranking algorithms clearly do a better job on one type of document or another. Some may adjust for word position or proximity of query terms. Others are partial to long or short documents or tend to give priority to term frequency instead of to term rarity in the **database** . Some may emphasize metadata; others ignore controlled vocabulary terms entirely. These are all reasonable design choices that may conform to a particular type of collection. While **searchers** cannot always understand why one **search engine** misses certain documents that another retrieves, we know that this happens. The differences in **search** algorithms may offer one explanation.

104

Evidence combination can refer to **searching** the same collection with different **search** engines and combining the results, or it can refer to using different sources to gather information about documents. For example, a collection of newscasts might be **searched** from speech text created by speech-recognition software. Closed-caption broadcasts would supply another source, and so would the video images themselves using image-recognition software. Each one of these sources is not a reliable source by itself - none of them contains enough accurate information on the subject of the document - but combined, the strengths of one make up for the weaknesses of another. Informedia (http:// http://www.informedia.cs.cmu.edu/), one of the first National Digital Library Projects, offers a good example of this technique.

105

Speech Recognition for Spoken Interfaces

Although we have become reasonably comfortable interacting with the computer by keyboard and mouse, it is not natural. Our interactions show it. Who would ask a spoken question with a single word? Yet, the vast majority of queries on **Web search** engines are single words. And would we really choose to input a query with parentheses and truncation symbols, given a simpler alternative? Spoken interactions are a more normal mode and a voice interface, or VUI (voice user interface), may solve some of the input problems that designers face with written or graphic interfaces.

106

There are two distinct sides to voice recognition: input and output. Speech recognition can go from text to speech or from speech to text (speech synthesis). Both speech recognition and speech generation software must be developed in order to create good VUIs. The easier of these is speech generation. People already can understand computer-generated speech because they already know how to adjust to slight variations in pronunciation or intonation, if only from listening to real people speak. Companies like Cogentex have already created technologies that generate speech from data plus a template. The Montreal weather **report** uses this product.

Voice recognition is a more difficult proposition. Natural-language processing gets us part of the way to voice-recognition systems, but a few levels of language important in speech make problems in written language.

107

The way we pronounce words has many more variations than we realize. For instance, the "c" in "cat" differs from the "c" in "core." Intonation patterns convey meaning by the song that is sung. A declarative sentence versus a question, for instance, is solely distinguished by the notes that the voice uses - a falling instead of a rising inflection. Voice recognition also stumbles on regional pronunciation differences, as well as on finding the boundaries between words. We run one word into another and expect our listeners to make the cut between each one. Computers can't manage this as easily. Try saying, "What's to stop me?" in a normal tone to see what I mean. Gotcha!

Nevertheless, voice interfaces have begun to appear. MyTalk (http://www.mytalk.com) from General Magic will fetch your e-mail and read it to you on the phone. It uses speech generation software and intelligent agents to read only what you want. You can interact using several hundred commands and, if you forget what to ask, it will give you choices. 108

Microsoft, with its SAPI standard (Speech Application Programming Interface) Persona Project, and associated Speech Recognition research groups, seems to be creating a successor to Microsoft Bob, which can interpret continuous speech and then generate an answer. Microsoft uses NLP and could apply this software to information retrieval as well. Other major players are Lernout and Hauspie, Dragon Software, IBM, Nuance, Motorola, Unisys, Dialogic, and AT&T. 109

Most of the research with NLP and speech recognition concentrates on understanding word boundaries and correctly identifying phonemes across diverse speakers and accents. This is a non-trivial task. One solution is to train a system within a small domain, such as answering customer-service questions for one particular company. Another is to train an application to recognize only one user's voice. This latter application is in demand for those who can't read a screen or type. In fact, reporters with carpal tunnel syndrome form a growing group of VUI users. 110

Once we solve the problem of establishing normal speech interaction as a computer interface, our whole mode of operation with computers will change. We will ask our car for directions and have it tell us where to turn next, after it has mapped out our route. In fact, that is available now. We will tell our agent to read us any news on **Internet** -related subjects while we make the coffee. It will ask us if we want to hear the urgent message from our boss first. And, we will ask for the monthly **report** to be generated from statistics and then presented as a PowerPoint presentation, complete with pie charts, without having to remember how to import a chart and resize it. 111

It's nearly 2001. Can HAL be far away?

Designing the Answer Machine 112

* Study users and their needs.
* Design the output and then the access system.
* Design the input to create the output you need.

Know your users and their work situation. Are they a captive audience? Can you offer training and will they take it? What kinds of information do they need and in what formats? Do they need in depth analyses, research reports, top-level summaries, weekly briefings?

Researchers look for information differently from **marketing** people or executives. This isn't surprising, since they all have different kinds of information needs. Researchers want in-depth information. **Marketing** people may want facts, statistics, or to keep up with the competition. Executives may want quick overviews and summaries that give them a lot of information at a high level in a capsulized form. Do your users want everything on a topic (high recall) or just the few best nuggets high precision)? How will they use the information? Do they need 24 hour, 7-day-a-week access from remote locations? What should the system output look like? 113

First, find out what your users need, want and how they will use the information. Then, design an access system that fits how they think and work. For instance, we have never found a user population that can distinguish between "subject headings" and "keywords." Don't expect that they will learn. Just create a system that doesn't require too much knowledge unrelated to their day jobs. 114

Create access models--subject, author, fields -- that make sense to your organization, even if this goes against library orthodoxy. If you only have computer science materials, don t expect that the Library of Congress 115

Classification will be useful. Think about why classification schemes were invented and then use something that can help distinguish among the materials.

Last, design the system so that it will give you what you have already specified. Don't get talked out of important features. And don't let fancy bells and whistles that will confuse the users to creep in. Keep it simple. Make sure that it is easy to navigate. Test it and retest it. 116

I am not necessarily a fan of all things automatic. The best systems give users an opportunity to interfere, add information, alter directions, and make corrections. These systems form a partnership with the user. When designing an information system, include the user in the design In the best of all possible worlds, system designers would observe how people use information within the work place and then design a system that fits into the normal work flow. 117

Conclusion

All these technologies add up to a seamless suite of information tools that will find information, organize it, keep it up to data, forage for patterns, and present understanding. In other words, an Answer Machine. The tools I have just described will enable us to understand large and complex sets of information more easily. These tools enable quick understanding by adding a new dimension of analysis and even fun to working with information. They will give knowledge workers the ability to examine, manipulate, and understand the information we retrieve for them. Using these tools, we can move up a level of abstraction to analyzing evaluating and planning. This will offer our profession an exciting, challenging role bright with promise. 118

To be involved in the development of the next generation of information system, we must be willing to think big, stepping back occasionally from deadlines and from gathering isolated facts and statistics. We must comprehend and clarify the place of information in the organization. This is a role for practical visionaries. 119

Fortunately for us, that's exactly who we are. 120

Susan Feldman is president of Datasearch, and of Datasearch Labs, a usability testing company for information products. She writes frequently on new information technologies, and tests, evaluates and recommends products for clients. 121

 122

Here are some samples of questions difficult to answer in traditional information retrieval systems:

Identify bacteria in the process of becoming drug resistant.

Identify Bermuda advertising campaigns that promote the island as a tourist attraction.

Provide articles and case studies on attitudes of companies towards media relations, including best practices for approaching the media and trends in media relations.

Provide information on "issues preparedness" (i.e., rationales for why companies should be prepared to manage a crisis or issue in advance and how companies can effectively manage a crisis or issue).

Provide information on "thought" retreats/seminars/ **executive** meetings, CEO retreats; and customer **entertainment** / appreciation events.

Identify books or articles that discuss how artworks through the ages have represented oral hygiene and dentistry (for example, is there a reason why the Mona Lisa doesn't smile?!).

Identify emerging competitors in X industry.

Where should I go for my vacation in January if I don't want to spend more than $600 per person and I don't like crowds? I'd like to go some place warm with nice scenery, somewhere near an ocean.

How many widgets will Zambia manufacture in the next 5. years? I just want a number for each year, not a pile of documents. I need this in a half-hour, by the way.

I need to keep upon new information technologies as they appear. (This means that I need to identify new terms and also to drop those that have become outdated.)

Tell me when my competitors have come out with a new product. I don't want any other press, releases.

WHAT EXACTLY DO MEAN BY MEANING?

People extract meaning from text on many levels:

* Phonetic is the actual sounds made when we pronounce words. This isn't pertinent to written text, but it does convey extremely important shades of meaning in speech.

* Morphological is the smallest unit of language which conveys meaning. This includes plural versus singular forms, as well as other prefixes and suffixes, like pre- or -ization.

* Syntactic is the role each word plays in a sentence. Many of today's **search** engines can parse a sentence, as we learned to do in elementary school, in order to pick out the subjects, verbs, objects, and phrases. This enables the engines to distinguish between Bill picked Al and Al picked Bill.

* Semantic is the dictionary meaning of a word, as well as the meaning of a word supplied by its context in the text. This level helps us to distinguish the difference in meanings of "pool" in "Let's play pool" and "Let's swim in the pool." This ability to distinguish among the many senses of the same word is called disambiguation. It enables a system to eliminate false drops. An NLP system should never give you financial institutions if you ask for erosion of river banks, not even for the Consolidated Bank of Moose River.

* Discourse is the structure of a whole document. Many documents have a predictable structure, such conclusion. Where a sentence is placed in this structure influences its meaning and its importance.

* Pragmatic is knowledge of the real word. For instance, when we say Europe, we know this geographic region includes France, even if the document never explicitly states this fact. This kind of knowledge can be added to new information systems so that the systems understand that Congressman Schumer and Senator Schumer are the same person.

(For a more extensive discussion of NLP, see Sue Feldman's article, "NLP Meets the Jabberwocky," http://www.online.com/onlinemag/OL1999/feldman5.htm, Online, May 1999.)

BEFORE AND AFTER

Smith Widget, Inc., October 5, 1999. 10:00 AM

Boss: Good Morning, Dennis. We need to update our competitive intelligence **report** today I'd like to know all the new products our competitors have come out with in the last 6 months, as well as any plans they have for new products.

Dennis: Okay. When do you need it? The president just called me and wants figures for the board meeting by noon today.

Boss: Well, I really did want it by noon too. It's also for tomorrow's board meeting. See what you can do.

Dennis: I'll do my best. What information do you need the most? I'll work on that first.

Boss: Well, I really need a list of new products and their **sales** figures listed by company, and then I want a summary of trends and predictions for the industry, just bullet points.

Dennis: I think I can get the list of products for you, since we already know. the names of the companies, and I've been keeping a file of the changes to their **Web** sites. At least we have the new product announcements. It'll take a while to wade through the documents I get from an online **search** though, so I'm not sure I can get you the other information right away. I'll do my best.

Boss: I really need them in a hurry so we can get the graphics people to turn them into a slide briefing.

Dennis: When is the board meeting?

Boss: Tomorrow at 1 PM.

Dennis: I can probably get you the bullet points tonight and give them to graphics for tomorrow morning.

Boss: Well, if that's the best you can do, I guess we'll just have to settle for it, but I did want to review the notes tonight.

Dennis: I'll see if I can give you some preliminary results by 5 today and then work on a summary and bullet points. We can give the **sales** figures to graphics as soon as I get them. They're the easier part.

Boss: Okay. Just let me know as soon as you have something.

Dennis: Okay. (Boss leaves, Dennis dials wife's office). Hi. Guess

what? It's quarterly panic time again. He wants a **report** by tonight. Can you call the Groves and ask if we can reschedule that dinner? No, I don't know the number of a good divorce lawyer, and I'm going to have a long enough day without any sarcasm. You know I love you. Sure, honey, see you when I see you.

Dennis (musing): Now where did I **store** that CI **search** strategy? Okay here's Dialog, here's NEXIS, here's Dow Jones. I'd better update the **Web** filter, too, and look at those documents in my widgets CI mailbox. Here are the strategies. Dialog, file 16:ss (Jones or Franklin or Thomas or Automated) (w) Widget? and (ec=65? or ec=33?)

Search 2:ss pc= and (ec=1? or ec=6?) and (predict? or projecting or projected or future or forecast? or trend or outlook or year0 (200? or 201?)

(2:30 that day)

Dennis (calling boss): Hi, I have the product info and **sales** figures for your three competitors. Shall I send them to you electronically? There were 467 documents from the online **search** , and I'll scan them as fast as I can to get you the info you need. I'm using Puffin **Search** to merge and relevance rank the **searches** I did in Dialog, NEXIS, and Dow Jones. Is it okay with you if I just start with the top 150?

Boss: Yes, but please try to scan the rest too. We really got into trouble when we missed that new company, Automated Widgets, last time. I think they are marginal, but it doesn't hurt to see what they're up to.

Dennis: I'll do my best, but the last train leaves at 9:30, and I have to catch it.

Boss: Well, give me what you have by 9:00.

OUTCOME: Dennis had to quit, having missed both lunch and dinner, at document 322, in order to have time to write the summaries and bullet points in time. Document 463 showed that Automated Widgets had hired an expert in networked appliances from Sun Microsystems. Smith Widgets was bought out by Automated Widgets in 2003. Boss took early retirement. Dennis went on to help create a company-wide information system, designing templates for interaction and categories for automatic indexing.

Automated Widget Company, October 5, 2009. 10:00 AM

Boss: Good Morning, Alvin. We need to update our competitive intelligence **report** today I'd like to know all the new products our competitors have come out with in the last 6 months, and any plans they have for new products.

Alvin, the Computer: Okay, boss. Do you want products from your competitors if they are in a different product category from Automated Widgets?

Boss: Yes.

Alvin: When do you want this? What format?

Boss: I need it by noon today. Give me lists of products, organized by name of company. Then I'd like of summary of trends in the industry. Just summarize and make some bullet points, but keep the information. I may want more details on some of the major points in the summary. We're really worried about MS Widgets, so give me everything you can find on them. I want recent hires and firings, and any industry analyst reports.

Alvin: Do you want **sales** figures like the last **report** ?

Boss: Oh, yes. I want **sales** figures for each. Compare them to the figures we have for that company 6 months ago. Just pull the old chart out of the last **report** and add a column for the new products and another for the **sales** . Also, give me any growth or decline in overall **sales** for each company. Don't forget their previous products.

Alvin: Anything else?

Boss: Yes, After you get me the lists, and the bullet points, update that competitive intelligence **report** we did 6 months ago.

Alvin: Same format?

Boss: Yes, but make the charts a larger font size. Also, extract the major points and put them and the charts in a slide presentation. Give me a separate slide on MS Widgets. I want that one by 1 PM.

Alvin: Anything else?

Boss: No, that's it.

Alvin: Okay. I will find lists of Automated Widgets competitors and their new products with **sales** figures and produce a list for each company. Then I will find trends and predictions. I will extract major points that appear in two or more articles or are mentioned several times

in one article. I will deliver these lists and bullet points by noon to your inbox.

I will update the CI **report** from March 31, 2009, and use the new major points and charts for a slide presentation. This can be ready by 1 PM, but cannot be printed by then. The **marketing** department has the color printer reserved all afternoon. Can you review the slides online, or should I notify the printer that your work takes priority? We can print after 4 PM.

Boss: I will review online. Make the print font big enough to read.

Alvin: 14 point type font?

Boss: Okay.

(11:30. Boss walks into room)

Boss: Alvin, is the **report** ready?

Alvin: It is ready, boss. Printed copy is in your inbox. Online copy is in the high-priority info box labeled competitive intelligence. Do you want me to read it to you or do you prefer to view it?

Boss: Read me the new products and major bullet points. Also anything you found that doesn't fit a category

Alvin: In the new product category;

Franklin widgets Programmablerefrigerator/stove module

MS Widgets Programmable bathtub module

Widgetech Programmable gas grill

Programmable clutter hider

In the people category, Andrew Wyatt gave a talk in September at the Futuretech conference. I summarized it for you. You met him at the WIA conference last spring, and I have a note to tell you to contact him in October. His phone number is 577-304-8976. His e-mail is aww@futuristics.com. I have his street address, too.

In the unpleasant surprises category, you didn't ask to monitorSolutions.com. It is a new company that matches your competitive profile. They have developed a "company's coming" remote control module that hides clutter, inventories the refrigerator, orders groceries, cleans the house, turns on the oven, and changes the sheets.

In the MS Widgets **report** , their earnings have gone up 23 percent. They have just acquired a widget integrating company.

Is there anything else you want?

Boss: Yes! Get me everything you can on widget integration companies. I want a list of those with actual products, and what those products are. Also, **sales** and predictions for each of them.

Add Solutions.com to our monitoring list.

Alvin: Okay boss.

OUTCOME: Automated Widgets is slugging it out with MS Widgets at the moment. Will either of them notice Solutions.com sneaking up on them? This is a case of dueling information systems. Winner take all. Which one has the better technology for raising red flags? Which do you think?

IMPLICATIONS FOR INFORMATION PROFESSIONALS

The dawn of a new era can be exciting or unsettling. Right now, there are so many fingers in what used to be our information pie that we may reel crowded and, perhaps, threatened. Computer scientists, psychologists, graphic designers, linguists, and **Internet** businesses are all carving out pieces for themselves.

What do we information professionals have to offer of value? First, we have a unique perspective about information itself. We understand how to ask the right questions in order to kind what we need. We understand balance in collections, good sources, and how to categorize materials so people can find them. This is invaluable. We also have something the others may lack -- we use information systems. We have **searched** for information for decades. We have practical experience. If we can temper the experience with the flexibility to try something new, we can become the part of the development team most firmly anchored in reality.

Things brings me to some tentative ideas on what to look for as you go about putting together an intranet or information system for an organization. These thoughts are tentative because they haven't been tested, and may theories are just as suspect as anyone else's. I can only rely on my own experience and tests of technology. Based on my comparisons of NLP systems with other systems, I know that NLP systems work and work well. Similarly, I have been extremely pleased with the agent systems and automatic indexing systems with which I have experimented. So, I know that

the foundation technologies work and much better than anything else I've tried. I think that if I we were putting together a system for tomorrow, though that I would look for products with these technologies as my base.

PUBLISHER NAME: Information Today, Inc.
EVENT NAMES: *360 (Services information)
GEOGRAPHIC NAMES: *1USA (United States)
PRODUCT NAMES: 7375000 (**Database** Providers); 4811520 (Online Services); 7372421 (DBMS); 3573025 (Document Processing Computer Systems)
INDUSTRY NAMES: LIB (Library and Information Science)
NAICS CODES: 514191 (On-Line Information Services); 51121 (Software Publishers); 334111 (Electronic Computer Manufacturing)
SPECIAL FEATURES: LOB
?

6/9/3 (Item 1 from file: 16)
DIALOG(R)File 16:Gale Group PROMT(R)

08059036 Supplier Number: 66217096 (THIS IS THE FULLTEXT)
The "R" Technology Revolution: Relationships, Research, Revenue.(Industry Trend or Event)
Arnold, Stephen E.; Colson, Michael
Searcher, v8, n9, p36
Oct, 2000
ISSN: 1070-4795
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Professional
Word Count: 9404
TEXT:
Single letters are the **marketer** 's touchstones. We have the ubiquitous "e" used in company names (eGain, an e-mail company), e-commerce (B2C and B2B varieties), and, as one pundit exclaimed, "E-nough." We want to focus on "R" -- relationship technologies gathering research to produce revenue.

Web or intranet **searching** is shifting from presenting a list of "hits" to a richer, more complex presentation of information, delivering information in a meaningful context. **Searching** is a modern technological wonder, but algorithms can only do algorithmic functions. Humans perform other types of functions, including putting information into context and drawing relationships. Combining the best of routinized spider indexing functions, adding a soupcon of statistical analysis, getting people back into the **search** process -- that is the future.

Searching is becoming more concerned with relationships of many types. It is becoming a regular task. When a **searcher** needs help, there are different types of intermediaries on hand. The best research combines relationships among data, other researchers, and various technological gizmos. The Napster-Gnutella-Freenet revolution has just begun.

It appears that **search** functions are shifting toward a distributed **network** architecture (DNA). As a set of services that works to integrate resources, DNA is an open systems, client-server technology. DNA integrates information from any data acquisition system, simulator, information **database** , plant monitoring system, etc., and then facilitates the acquisition, archival, and retrieval of all types of information created or gathered. This model addresses the development, deployment, and maintenance of **Web** -based applications in an effort to build relationships among people and data.

The letter "R" may step out from the shadows of **searching** to the center stage. The themes that this article touches upon interconnect and directly relate to exploiting the "relationship" technology of the **Internet** :

* **Searching** and retrieving information accessible through a browser and connections to the **Internet** where other people, and their experience, are online at the same moment and accessible with a mouse click.

* Finding connections between and among information in different formats and types.

* Using **Web** sites that are magnetic (that is, pulling traffic in) and sticky (keeping users on a site and coming back to it).

Today's new technologies can solve "R problems"; that is, perform regular" tasks and attack "real" problems with the click of a mouse from virtually any location at any time. Most importantly, "R" sites generate revenue.

Searching : One of the Core **Web** Services.

Finding information is becoming a larger part of the average **Internet** user's daily life. **Searching** touches virtually all **Internet** users. Even locating information in one's electronic mail folders can be a painful experience. In Outlook Express it is difficult to sort mail from a particular sender by date. On most intranet systems, such as those from Plumtree or mynet.com, among others, pinpointing a particular news **story** requires patience as well. When one wants a particular chunk of information that resides on a particular organization's **Web** site, the **searcher** often finds the task time-consuming, tedious, and sometimes impractical.

Yahoo! (http://www.yahoo.com) is one of the most-visited sites on

the **Internet** . Like Lycos and FAST's services, Yahoo! has an international presence. The Yahoo! splash screen is not particularly user-friendly. Some might consider it cluttered (see Figure 1 on page 391.

Two points, however:

The Yahoo! visitor can locate information by pointing and clicking. No instruction other than the basics of pushing a mouse button is required.

Special services and features pop off the page. Good use of color and icons catch the user's attention and invite a click.

Search -and-retrieval experts often ignore the Yahoo! **search** function. To a great extent, real live people create the basic listing of sites arrayed in a taxonomy or classification system in Yahoo!. Many on the Yahoo! editorial teams overseeing the 400,000 or so taxonomic entries have degrees in library or information science. However, Yahoo! really only directly indexes certain sites, now relying on Google to handle the rest.

When a user clicks on an entry in the taxonomy, a list of sites appears. These listings have been created by a Yahoo! professional or by a person who has submitted a site to Yahoo! Even the submitted listings that appear in the taxonomy have been reviewed and possibly edited by the Yahoo! editorial team.

At the bottom of the Yahoo! taxonomy pages links appear that invite the user to review a list of the 101 most useful **Internet** sites. A click triggers an advertisement for the magazine, **Internet** Life. For users who want to explore the news, Yahoo! offers a point-and-click approach plus an advanced **search** function. Based on information provided to our researchers, however, fewer than 10 percent of Yahoo! news users access the advanced **searching** function (see Figure 2 on page 39).

The quest for better **search** -and-retrieval systems continues. AIT's **Web** site (http://www.arnoldit.com) provides links to more than 600 **Internet** directories and **search** engines that index content outside of North America. Within North America, the number of **search** engines and directories runs into the thousands. These are some of the more interesting new services:

* i411.com--a service that provides high-speed directory **searches** for telephone listings, various types of fact-based reference products, and other types of look-up information.

* Vicinity.com -- a service that allows a user with a mobile telephone to "aim" the mobile in a direction. The service then displays restaurants and theaters along that vector.

* Prompt Software -- a **Web** -based **search** -and-retrieval engine that displays hits, extracts keywords and phrases, and generates an abstract or synopsis of a site's content when the user "hovers" a cursor over a URL.

* Ixquick.com -- a metacrawler service that provides a simultaneous **search** of multiple sites. When results display, a score for the importance of the site is calculated using a combination of popularity (how many "clicks" or "hits" a site receives in a time interval) and link analysis (how many other sites point to that site).

* Napster, Gnutella, and Freenet -- these are "sites" that provide users with a directory space that they can **search** . Napster (http://www.napster.com) carries information on music that an individual has posted on his machine for others to download. More interesting is Wrapters (http://notoctavian.tripod.com), which allows a user to swap non-MP3 files over the Napster system. iMesh (http://www.imesh.com) offers Napster-like features and can be used to find digitized movies.

The emergence of "person-centric" services represents one of the more interesting innovations in **search** and retrieval. The Arnold IT **Web** site (http://www.arnoldit.com) lists more than a dozen **Web** indexing and directory services that provide human-intermediated **search** help, such as Abuzz.com (http://www.abuzz.com), Askme.com (http://www.askme.com), and Keen.com. Other sites "sell" **search** help: for example, Frenzi.com, which uses a barter system and Exp.com (http://www.exp.com).

The **search** service that really breaks new ground is Napster, the embattled file-sharing site (see Figure 3 on page 39). The music industry has worked diligently to fundamentally alter the Napster service, but the technology of distributed content and virtual indexes is moving faster than the music industry's copyright lawyers. (A fierce legal battle swirls around Napster and its threat to copyright infringement. At one point, a federal district court ruling threatened to shut down the operation. By the

time you read this article, this may have occurred, but at presstime, the 9th U.S. Circuit Court of Appeals had established a three-judge panel to hear arguments in early October. The recording industry wants the service shut down, but a diverse array of groups -- ranging from the Consumer **Electronics** Association to the Association of American Physicians and Surgeons--has filed friend-of-the-court (amicus curiae) briefs on different aspects of the case.) Gnutella and Freenet offer similar functionality, with t he "virtual directory" distributed across different machines, not centralized in one location.

The Napster interface provides a glimpse into the functionality embedded in the concept of distributed **network** architectures, where the opposite of a centralized repository and index exists to satisfy requests for information:

Gnutella, Freenet, and other distributed or "virtual indices" will have a significant impact on **search** and retrieval over the next 2-3 years in " **Internet** time." The overhead associated with using a series of scripts like Inktomi, Northern Light, or AltaVista's Raging.com service to index the **Web** is enormous. When users do the indexing, core **network** functionality creates a "virtual directory" available for **searching** .

One new service, InfraSearch (see Figure 4 at right), uses the Napster distributed directory paradigm, combining both the technology and pricing mechanisms. The developers of InfraSearch came from the University of California-Berkeley's Experimental Computing Facility (XCF) and were responsible for Gnutella.

A publisher or information producer with content provides Napster-like information, substantially self-defined, to InfraSearch. A query returns a pointer to the file plus whatever meta-information that the submitter provided. InfraSearch can, with the information producer's permission, spider the content and automatically create the brief description and the pointer. Content providers can specify if the information is provided without a fee or carries a per-view surcharge. The beta version of the service provides access to content from Moreover.com and Yahoo! **Finance** , services that carry both free and for-fee information.

When a document is located, a click takes the user to the publisher's site where they can view the document for free or after paying a fee. Infrasearch uses the distributed **network** architecture in an interesting way that is the opposite of the "old" Dialog model.

As more and more PCs keep plugging into the **Internet** , the physical location of particular chunks of data that a user may wish to access has become increasingly irrelevant to **Internet** users. Many newcomers to personal computing have only foggy notions about directory structure, folders, file names, and the location of devices that actually hold the data or even the application software.

Inktomi and Google have transformed the spidering **business** into a **business** model similar to that of an Application Service Provider. Users "rent" or "license" an application or, in Inktomi's case, indexes of **Web** sites. When a person uses a **Web** **search** feature on a **Web** site such as Microsoft's or VerticalNet's, they do not use a **search** **engine** created or operated by that **Web** site. Instead, third parties have created the taxonomy and index. It is cheaper for Webmasters to "rent" a spidering utility and avoid the multi-million-dollar operational costs of **Web** indexing.

Back to People

For information professionals, however, the most interesting trend in **search** and retrieval is probably the emergence of person-intermediated services. The trend extends well beyond Yahoo!'s directories, largely built by staff editors or people submitting sites for inclusion in Yahoo! Table 1 on page 38 provides a roundup of some of the **searching** services that involve humans in the **search** and retrieval process. Note that the approach varies from volunteers (Mozilla open directory) to for-fee services (Guru.net), with numerous twists.

Unlike About.com (formerly the Mining Company), expert services are shifting from **search** and retrieval that **stores** answers to anticipated questions to services driven by the asking of questions. The **market** leader in "answering questions" is Ask Jeeves (http://www.askjeeves.com).

The premise of About.com (see Figure 5 below) is that some individuals gather and maintain links to **Web** sites on particular topics.

About.com's resident expert on **Web search** and retrieval, for example, is Chris Sherman, an information broker.

How does Mr. Sherman's listing of **Web search** sites differ from "hits" produced by a spider-generated service like FAST (http://www.alltheweb.com)? Presumably Mr. Sherman uses other sites' indexes (probably ones built by spiders) to assist him in his research, so About.com's model is closer to Snap's or Yahoo!'s editor-built directory. But in this case, Mr. Sherman is the person responsible for what appears in his list of links. It is quite difficult to find out what person or persons is directly responsible for a section of links at an **Internet** editorial shop like NBCi, owner of the Snap and Xoom directories, on the other hand.

How do the listings differ? First, Mr. Sherman constructed his core list using other research tools. The sites listed are those that he judges the most useful and pertinent to the subject of **Web** research. Second, if a person scanning Mr. Sherman's links wishes to grouse or provide some other type of feedback, Mr. Sherman is available to answer questions by electronic mail or by telephone if the matter is urgent. Third, the links are updated on a cycle set by Mr. Sherman, not a script. About.com is a quite useful service, particularly for a person who wants to look at focused hits on a topic.

Ask Jeeves (see Figure 6 on page 40) is the natural language champ, at least in **Internet** space and stock trader visibility. For the month of June 2000, Media Metrix (http://www.mediametrix.com) noted that Ask Jeeves had climbed to the number 15 spot in unique **Internet** visitors. For that month, Ask Jeeves had about 12 million unique visitors -- or about 400,000 a day. This compares to America Online's 2.6 million unique visitors per day. AOL was the most visited site in June 2000. Ask Jeeves is expected to lose about $40 million in the quarter ending June 30, 2000. The stock soared to 190 per share and sagged to 18 per share by the end of June 2000.

Experts in NLP scoff at Ask Jeeves. The user does enter a natural language query and does get an "answer." Unlike the more esoteric NLP engines crafted at Carnegie-Mellon or Syracuse universities, Ask Jeeves looks at a question, like "What is the weather in Capetown, South Africa?," and matches the question to a master list of 10,000 or so questions that Ask Jeeves "knows" how to answer by hitting a **database** of URLs and scripts. The user sees a template that allows them to select from a suite of likely "answers."

So What's Next?

To bring this quick overview of **Web search** -and-retrieval to a close (see Table 2 on page 43), one thread unites these quite different examples. Each of these sites makes a concerted effort to put information into a context of some type. Yahoo! uses a very easy-to-grasp taxonomy or simplified list of Library of Congress subject term headings as the centerpiece of its site. This core function is surrounded by mail, chat, discussion groups, and dozens of other functions that allow the user to look for information in an interactive, text-based environment. For the handful of users who want to formulate a precise query, Yahoo! has crafted a form to walk the user through the construction of a Boolean query. But most users do not use this feature. The point-and-click approach allows the user to spot something of interest, click on it, and explore other related topics in Yahoo!'s presentation of related links.

Consider Twirlix (see Figure 7 on page 40). When a user does a query, Twirlix runs a query over its index of **Web** sites. Each hit display carries a thumbnail picture of the **Web** site plus a brief textual description of the site. The company calls this "Quick Preview." Twirlix's "TV-Preview option" presents a list of sites with Quick Previews, QualityRatings, language, and site name -- but presented consecutively from left to right in order to get maximum information on a single page.

As noted, Napster, Gnutella, Freenet, Infrasearch, and the other "virtual directory" **searching** services make users and their information needs the focal point of the service. A **searcher** has a direct relationship with another user who has created the index pointer and created the file to which the index entry points. In effect, these new distributed services could disintermediate other **Internet** middlemen from the **search** -and-retrieval process. This is unlikely, however, because people want vetted lists of sites. Napster has, like Yahoo! and Lycos, created a host of what might be called "regular" communication tools. There are bulletin boards, chat groups, communities, and similar services to

allow the users of distributed services to talk and meet one another on common ground.

About.com and a number of person-centric **search** and research sites make an expert the center of focus. Questions, **searches** , dialogs, and other interactions are encouraged by these sites. The researcher is in close relationship to the person responsible for the links in the site and can, with a click or two, engage that person in one-to-one conversation.

Ask Jeeves allows the user to ask the system a question. The answer, however, is placed in a context. The user can scan the presentation of different ways to get the answer to a particular question. Like Yahoo!, Ask Jeeves presents a context-rich array of choices. Recently Ask Jeeves has followed Yahoo! in offering personalization services and other types of communication vehicles.

The "relationship" technologies or techniques in this handful of high-traffic sites follow a common theme:

* Information is put in a context. Users can point and click on links that they deem the most interesting, most useful, or most promising of the choices presented. A new or experienced user can scan what is offered and make a choice using the other Links as points of reference.

* The human-built sites make it clear that an individual or group of individuals made choices about what sites to include. These researchers have an understanding of their topic area and have decided what to include in the list, how to explain the site, and when to update. The list of links is constructed on the premise that an informed person has a close involvement (relationship) with the content presented. Someone, not some "thing," is responsible.

* What might be called "regular" communication tools abound. The researcher can ask a person a question and get an answer by electronic mail. Bulletin boards and chat groups allow a researcher to ask others in a public forum for help.

What About NLP?

NLP or Natural Language Processing remains a popular research topic among Ph.D. candidates in information science. The Holy Grail they seek would have someone speaking a query into a mobile device and the **search engine** understanding the query, collecting results, and delivering them to the user.

NLP is computationally intensive, fiendishly expensive, and far enough in the future that only the most ardent optimists bet that an NLP **search engine** will unseat Yahoo!. Ask Jeeves is not NLP-based on sophisticated linguistic and statistical algorithms. Ask Jeeves matches a query to a collection of canned answers. When a match is found, the **search** is run. The company is upgrading its technology, but its importance in the development of **Internet searching** lies in its ability -- or attempt -- to offer a service that allows a user to enter a query and get a response ma second or two. True NLP engines are best demonstrated in computer research laboratories.

A study by NPD New Media Services (http://www.npd.com) revealed what information professionals have known for many years and what information scientists have verified with hundreds of Journal of ASIS articles: An **Internet search** returns "better results" when the user relies on "multiple-keyword **searches** and the use of more than one **search engine** ."

However, based on the study of 33,000 **Internet** users picked at random during the first quarter of 2000, 81 percent of **Web searchers** reported "success finding the information they were looking for." The sponsors of the study were 13 **Internet search** engines, including AltaVista, America Online, Ask Jeeves, Excite, Go, Google, GoTo.com, Hot-Bot, Lycos, MSN **Search** , Netscape **Search** , **Web** Crawler, and Yahoo!.

Another interesting finding: Nearly 45 percent of **Web** users **search** using multiple keywords. About 29 percent rely on a single word. Only about 18 percent of those surveyed formulate queries in the form of questions. Nearly 80 percent use the same keywords in different **search** engines. Only 19 percent reformulate their query. (To view the survey results released on July 17, 2000, go to http://www.npd.com or http://www.iprospect.com.)

The company's research summary noted that a " **Web** site must rank in the top 30 results to get significant traffic. Most people do not scroll past the top 30."

The practical approach to NLP is to factor humans into the **search** equation. After all, even the least bright human can figure out some of the nuances that bedevil the use of spoken language to derive meaning from context, intonation, and implication. **Search** engines are finding out that humans are cheaper, better, and faster than most NLP engines available in the **marketplace** .

The "R" revolution can be seen in the use of people, information in context, and communication tools. The technology enabling these functions is remarkable. The use to which the technologies are put are easy to understand. The consumerization of the **Internet** has changed the **Internet** from a narrow communications channel to a multifaceted one.

Metasearch Engines

Metasearch services allow users to enter a single query, which is then passd on to a number of **search** engines. Some services, like Ixquick, allow the user to specify the engines to which the query is sent. Other services, like Copernic and Bull's Eye, provide preselected lists of **search** engines for particular types of queries. One bundle of **search** engines works for news **searches** ; another set for **Web** **searches** , and so on. Bull's Eye offers more than 60 selections of **search** engines, giving the user fine-grained control of the metasearch function.

The popularity of technology that takes one query and runs it across multiple **Web** indexes stems from one painful fact. The spiders and humans building **Web** indexes cannot keep pace with the amount of new information posted to the **Web** each day. Furthermore, a larger and larger percentage of **Web** pages are dynamic and exist only when a user takes an action. Agents and spiders can be programmed to create pages from dynamic content, but it is expensive to create script libraries. As a result, **Web** indexes focus on the popular pages. A metacrawler plays the odds that most **Web** indexes have less than 30 percent overlap. The result is that the user takes advantage of the lack of duplication. The results typically appear in a list of results ranked by relevance. Thus, instead of looking at separate lists of hits, the metacrawlers provide some relationship among the results.

Seymour Rubenstein, the creator of the ground-breaking WordStar word processing software and Quattro Pro spreadsheet, has developed a remarkable metasearch engine for his new company, Prompt Software (see Figure 8 on page 42). In addition to querying dozens of **Web** indices, the software can be aimed at intranet servers and special types of documents in Word or PowerPoint format. In addition, the software includes a feature that allows the user to see an abstract or summary of the information on a **Web** page simply by holding the cursor on a site's hyperlink. But the most remarkable innovation is Mr. Rubenstein's algorithm for extracting a word list with proper names and compound words. A user can scan the alphabetical list, click on the precise term, and be launched to the **Web** page where that term or phrase appears.

Popularity Engines

Popularity algorithms are used to determine what to index and relevance. Stanford-incubated Google is the pioneer in link analysis. A **search** on Google returns results based on sites that have links to them. The idea is that the more important a site, the more other sites will point to it -- the better the "seed," the taller the tree.

An alternative approach is to count the clicks. Direct Hit, hatched at the Massachusetts Institute of Technology, indexes sites that people use. The idea is that popular, important sites are used most often.

Blending the two ideas is Ixquick (see Figure 9 on page 42), based in New York City. This site combines link analysis and click-stream analysis. Sites that rank highly with both "score" higher in importance than those that do not. Ixquick is interesting because it focuses on the relationships between links, clicks, and the user's query across text, MP3, and image files.

What High-Traffic Sites Have Discovered about **Searching**

Demographics. It has been discovered that if a site does not get traffic from people in the "right" demographic, it is difficult to demonstrate its success.

No one knows how many sites exist "on" the **Internet** . The number is somewhere in the millions, with an estimated billion or so pages. Estimates for the volume of data flowing into the **Internet** are equally fuzzy. Received wisdom argues for the **Internet** 's doubling every 3-6 months. Pick

a number. The point is that there are hundreds of millions of users, sites, pages, or any other metric one wants to use to measure the **Internet** .

A Dartmouth College **report** claimed that **search** engines are struggling to keep up with the new content and changes to existing **Web** pages. According to the Dartmouth researchers, "We were able to determine that one in five **Web** pages is 12 days old or younger." Only one in four pages is more than 1 year old. Not surprisingly, the top sites include a **search** function, but the days of pure **search** creating a Top 20 site have been left behind. **Searching** is not enough. (For a copy of George Cybenko and Brian E. Brewington's paper, "How Dynamic Is the **Web** ?," go to http://www9.org/w9cdrom/264/264.html.)

With an estimated 600 new pages added to the **Web** every minute, **search** engines have two difficult task to perform rapidly. The first is to index the new pages as soon as they appear. This means that the agents being used to monitor **Web** sites must operate continuously from multiple servers and multiple **network** access points. Identifying changed pages places additional demand on the spidering infrastructure. When a change is detected, the spiders must reindex the site. The combination of monitoring for new pages and determining changes to previously spidered and indexed pages requires substantial amounts of money and technical talent.

The surprising aspect of the **Internet** is that it is behaving in a decidedly old-fashioned way when it comes to user behavior. Technology wizards may stay up all night cursing America Online, but AOL is doing something right. It makes money. It has customers. It is one of the top sites on the **Internet** . A look at one of the popularity contests or hit parade of **Internet** success **stories** (see Table 3 on page 51) reveals the disparity between a typical company site with a few hundred to a few thousand hits per day to the top-performing **Internet** destinations (formerly called portals and now positioned as networks).

An inventory of the "R" technologies offered on these sites reveals the following:

* Every top-rated site offers one or more " **search** " services. Some, like Yahoo!, use a combination of services. A directory assembled by a skilled team of editors is supplemented with entries submitted by creators of sites supplemented with the spider-built index from Go ogle. If a Yahoo! directory entry does not suffice, a click launches the query against Google's index. If these services fail, Yahoo! offers a mind-dizzying range of clubs, discussion groups, and real-time communication options.

* Virtually all the leaders offer bulletin boards and/or discussion services. Posting, reading, and replying to messages related to a subject close to the heart of users is a ubiquitous "R" function.

* AOL, Microsoft, and Yahoo! offer real-time messaging. The three top services offer real-time services, while companies like Gooey provide alternative tools.

* Virtually all include some type of personalization features. Due to the cost of building intelligent personalization tools, many sites, including Yahoo!, opt for user-constructed "views" of information and services. As the cost of high-end, high-maintenance, distributed, multi-agent reasoning systems drops, personalization will become a standard feature.

* All had easily located, automated customer service mechanisms. With questions from users running well over 600,000 per month at the most-visited sites, automated customer service is a must.

* Virtually all tailor screens to the actions of the user or a user-completed profile so that "sell through" offers relate to user needs or behaviors.

The information in Table 3 reinforces the shift to what we call "Relationship Technology." The sites pull users; that is, they are magnetic. They hold onto users within the " **network** " of sites; so they are sticky. Research by Forrester, Gartner Group, Nua, and Giga Group, among others, has reported that the longer a user remains within a site, the more valuable that user is in terms of advertising and sell-through. (Sell-through connotes that a person makes a purchase that benefits the site operator while within the site. The purchase can be a referral, in which case the referring site gets a commission. See, eContent, August-September 2000, for a more detailed discussion of pricing, "The Joy of Six" by Stephen E. Arnold.)

First, look at the difference between the top five sites with an

average number of unique visitors in the 43 million range. The average number of visitors to the sites ranked 16 through 20 is about 11 million - one-fourth the traffic. The top five sites are not really sites. AOL, Microsoft, Yahoo!, Lycos, and Excite@Home are really full-scale online services offering **searching** , shopping, community, **business** information, and **entertainment** services.

Second, the shift from sites that provide one core service is profound. One year ago, sites that offered " **network** " services were limited to America Online and the much-maligned Microsoft **Network** (MSN). An **Internet** "site" provides a wide range of services. In fact, in order to capture "clicks" and unique visitors, these **Internet** sites are positioning themselves as "networks." Seven of the 20 sites use " **network** " in their name as part of the site's positioning.

Third, the best way to grow is to purchase other high-traffic sites and roll those users into the "mother ship" site. Sites that have followed this practice are Lycos, Excite, NBCi, America Online, eBay, and recently CNet and ZD Net. CNet bought ZD net in a staggering billion-dollar deal.

If we look in the **business** -to- **business** space, a similar **story** is unfolding. The explosion of **business** exchanges in the chemical, steel, and medical supplies industries is following the trajectory of the "consumer" **Internet** segment. A good example of the blend of back-office services and relationship technologies can be found in the sites built by Time0, a unit of Perot Systems, which has a partnership with companies such as Grainger's OrderZone (see Figure 10 on page 42).

The features and functions of the **business** -to- **business** sites are similar to those found in the public " **network** " sites:

* One-click access to functions
* Supplementary information about products
* News and information
* Communications, including bulletin boards

The shift that has taken place in the last 24 months has been similar to what happens when a piece of tape is pressed against grains of sand. The tape picks up the separate grains and becomes something quite different from the original two materials. Sandpaper has characteristics, qualities, and applications quite separate from the "paper" and "grit" of the original compounds. Figure 11 on page 45 provides a simplified view of what the **Internet** 's ecology supports.

The **Internet** has functioned like adhesive tape since its inception 30 years ago. If we extend the sticky tape metaphor to a **network** infrastructure, the transformation of a redundant **network** into the global phenomenon becomes quite interesting. The original **Internet** performed the file transfer and machine access functions that its designers intended. But the **Internet** has proved to have powerful adhesive and adaptive qualities. Consider the concepts that appear in "Basic R Technology Drivers."

* Migrating. An excellent example is the move of voice telephony from its analog roots with a long history of getting more voices to move down first copper and later optical fiber, switching calls from one place to other places, and so on. Voiceover **Internet** Protocol or VOIP allows ordinary telephony, plus videoconferencing, with none of the six-figure set-ups required just a year or two ago. Voice chat, whiteboarding, and application sharing will benefit from migrating an analog service to the **Internet** 's digital ecology.

* Transplanting. The idea is for an entrepreneur to pick up one technology, application, or function where it was developed and may have thrived for years, even decades, and to move it to the **Internet** . Examples of transplanting include electronic mail, bulletin boards, calendars, greeting cards, and news headlines, among hundreds of other functions.

* Fusing. Technologies take two or more **Internet** services, technologies, or functions and glue them together. Examples range from Gnutella (open source software, file sharing, and persistent connections) to real-time personalization (cookies, agents, and scalable databases).

* Building. Engineers grab multiple functions and assemble them to build new subecologies. Examples of this type of innovation range from the Application Service Provider **business** to the emergence of comprehensive, global **markets** for such products as specialty chemicals. Using the **Internet** as a construction zone, engineers build full-scale service bureaus. These virtual back-offices perform insurance, banking, factoring, and other mundane **business** functions for clients who want to slash

operating costs and reduce time in operational cycles. With the **Internet** 's layered ecology, a technology or innovation can be put in several different "places." The revolution in distributed file and directory services is now driving Microsoft Corporation's company-wide realignment. File and directory services provide the lubricant for the turmoil swirling around unauthorized distribution of digitized music, videos, and text. The use of the **Internet** as a place where colleagues can meet to view a PowerPoint presentation and dis cuss a project in text chat or voice/video media is simply a different manifestation of **network** plumbing, operating system services, and applications.

Internet technology gives ideas and technologies the breeding drive of lonely gerbils. Neither gerbils nor the **Internet** is likely to cease breeding in the foreseeable future.

Bottom Line -- What Makes Winners?

We started with three "R's": research, relationships, and revenue, which boil down to three metaphors or connotations. Our interpretation of the developments that have been highlighted are somewhat new, and we invite comments and feedback.

Figure 12 on page 46 illustrates the four stages or steps that users go through before they decide to make a particular site their "favorite place." The first stage is access. Users have to be able to get to the site. Sites that respond slowly or put too many barriers in front of the user will probably not build a strong base of repeat users.

Once the user has access, the site must provide functions and services that allow the user to meet specific needs. The better the site is at meeting the needs that users may not be able to verbalize, the more likely users will return to the site and stay within it. Our research reveals that users under the age of 21 leave a site within 15 to 20 seconds of its splash page displaying.

The features that keep a site fresh place incredible financial and technical demands on the site's owner. Dynamic pages, personalization, and special features like Wireless Access Protocol (WAP) pages make static, text-based pages look stale, even when carrying valuable information. The site must provide ways for the users to interact. The site must also provide a broad, interesting selection of regularly refreshed services. AOL, Microsoft, and Yahoo!, among others, make it possible to refresh sites in near-real time. Such change keeps pages magnetic and enhances stickiness. Core communication functions make up the user's part of the site's experience. The messages and information users post become "homestead content." Users have a vested interest in the site.

The final stage for a successful site is enhancing the site's appeal as the only place to go for most services. America Online and Yahoo! are among the most successful nesting sites in the consumer space. Users literally do not know what to do if their AOL or Yahoo! page is not available for some reason.

Based on work sites large (US West Yellow Pages) and small (Talavara), a successful site follows at least four steps. Some can jump over all phases in a matter of weeks. An excellent example is the explosive growth and acceptance of the Napster music site. Other sites gather momentum over time, moving through evolutionary phases with almost predictable enhancements. Examples include the evolution of Amazon.com from a vendor of books to a Wal*Mart with "homestead content," **data mining** suggesting books and music to the customer, and a widening stream of commercial products on offer. eBay has moved along a similar path with a foray into guarantees, credit-card billing, and a print magazine.

Figure 12 is like a whirlpool sucking uses toward an "R." High traffic sites exert a suction of sorts. People send electronic mail to their friends about a useful site. Conference speakers call attention to sites that do something that can attract users and attention. "Buzz," the vibration from "viral **marketing** ," gets people to take a look. If the site has something that offers value to the user, the user will keep coming back. People, even those not interested in New Wave music, will take a look at a hot site once to see what the excitement is about. The result is a digital cyclone that pulls users, a "force field" that turns the site into an electromagnet pulling users into it.

The four "stages" or components of the diagram are certainly arbitrary. In order to have a successful site, people have to get access to it. America Online relies on "carpet bombing" prospects with almost

unavoidable offers of free access. Other sites like the Food **Network** 's pages about the Japanese cooking program Iron Chef happen without a budget and almost accidentally. Sites that offer day traders access to financial information from a high-profile source like Thomson Corporation charge for access. Sites that offer access for registration only generate revenue by selling names, setting up click-through and commission deals, or selling advertisements and sponsorships.

After they have people coming to their site, savvy site operators put text discussion groups in place. Others build systems that allow users to post comments on virtually any topic and then use **data mining** tools or clustering algorithms to provide views of this grassroots content. Discussions and posting software are available from many different sources, including PowWow, Buzz Power, and ASAP-ware, to name just three.

As the message traffic grows, the successful site enhances the interaction options for its users. The high-traffic sites create environments that **support** live conferencing, automatic profiling and filtering of new messages, voice, video, and a constantly expanding blend of services designed to get people interacting with other people. Among the most interesting examples are sites that blend **Internet** directories and electronic mail access with experts who compile a topic-specific directory. If a directory listing falls to help, the user can "talk" directly to a person who knows about the subject.

The goal is to get users or customers who come to a site to stay there. "Nesting" is the high goal of the site's technology, **marketing** , and value. Our work with high-traffic sites supports what is emerging as common sense. Customers who stay within a site are more valuable than a visitor who clicks once and is gone in 20 seconds. Nesting users have value as "eyeballs," potential purchasers of goods and services on offer, and as what is called, somewhat awkwardly "commitment." Site operators feel that a person who cares about a site will respond to surveys, provide information via online forms, and promote the site to others. The **marketing** information extracted from habitual users can provide important clues to the site operator about what features the core group of users and customers desire.

If we contrast the public Ask Jeeves page with the "new" personalized Ask Jeeves page (see Figure 13 above), three differences strike one immediately:

* Ask Jeeves has added news, one-click access to weather, and other types of information.

* The user is asked to register, providing Ask Jeeves with a way to monitor what the user does.

* The **search** function has moved from the primary purpose of the page to a secondary function.

* Persistent frames surround sites to which Ask Jeeves points. A user knows he is in the Ask Jeeves space and has a relationship with that site while still viewing information from another service with a different brand.

Web sites are changing in the consumer and **business marketspace** because operators have found single-function sites difficult to sustain. To illustrate, one year ago, of the sites listed in the top 20 sites table, only three were full-scale portals: America Online, Yahoo!, and Excite. One year later, 17 of these sites can be considered as "mini-AOLs," offering a wide array of information services, functions, and activities for users.

But does the same trend appear in the **business** to **business market** sector? Consider the screen from SageMaker's intranet toolkit (see Figure 14 on page 48).

The similarities include a visually striking presentation, a combination of exposed information and news and one-click access to branded sources, and a tabbed interface that facilitates location of work-related information.

The future, however, is rich media. In closing, check out **Finance** -Vision, the multimedia Yahoo! service.

Although the interface is composed on many different panels, four key changes from "standard" Yahoo! are evident:

* The box with the large "Y!" presents streaming video. **Finance** -Vision is designed to **support** rich media as a core function.

* Personalized Yahoo! appears in the panel on the left side of the screen. The user can control stock quotes, news, and other information

services that are updated a user-defined schedule.

* Normal browser functions operate with hot links between the information that accompanies the streaming media or with the option of ad hoc **searches** .

Why Relationship Technology Matters

Jeremy Rifkin singled out the "R" in his book Age of Access (The Age of Access: The New Culture of Hypercapitalism, Where All of Life Is a Paid-For Experience. New York: Jeremy P. Tarcher. Putnam. 2000. ISBN 1-58542-018-2). Mr. Rifkin calls attention to the importance of access -- for example, the disadvantage to a person of exclusion from an online community. His book points out that access is growing increasingly important.

The "R" in Mr. Rifkin's book stands for "relationships" in cyberspace and the everyday aspects of life. The technologies that facilitate relationships have helped high-traffic sites keep their users/customers coming back and staying in the site for longer periods.

The technologies that drive relationships vary. Sites offer free **Web** pages so "members" can share their ideas with others. Geocities was one of the first of the free **Web** page services. Thousands of sites offer this service now. The technology is far from trivial, but once the trail is blazed, following becomes routine. **Data - mining** services like those pioneered by Firefly Technologies (acquired by Microsoft), Net Perceptions, and the less well-known bluestreak (http://www.bluestreak.com) allow a site operator to do the following:

* Monitor user behavior in near-real time
* "Fuzzify" or generalize individual behaviors into categories or clusters of behavior
* Place individuals into a generalized cluster
* Identify what similar people buy or use
* Locate similar products and services

Suggest these products via banner advertisements or personalized electronic mail to other members in a cluster.

Relationship technologies that monitor behavior are subject to intense scrutiny from consumers. bluestreak (Newport, Rhode Island) has a product called Real-time Access to **Data** , **Analysis** , and Response, or RADAR. bluestreak's next-generation tracking tools preserve a user's identity. bluestreak's major customer is AT&T and its first application of its technology will focus on evaluating the performance of banner advertisements. bluestreak analyzes the standard **Web** log data generated automatically by Apache, iPlanet (Netscape), **Internet** Information Server, and others.

bluestreak has developed a proprietary suite of statistical and analytic tools that determines what is used on a site. Armed with that data, bluestreak predicts click-stream patterns. What is remarkable about blue streak is that its On-the-Fly technology does not use "cookies." These chunks of data that **Web** sites write to a user's computer have become increasingly controversial, particularly in Europe. Cookies allows a broad array of functions, including monitoring, without the user's knowledge.

Personalization is another major aspect of relationship technology. Yodlee (http://www.yodlee.com) is a developer of specialized personalization services that are being integrated into many sites. Yodlee's technology can build a customized **Web** page for a user to provide a control panel for that user's personal financial activities. Personalization services allow a user, customer, or member of a **Web** service to create a personalized "view" of information. Yodlee's customers include Intuit, AltaVista, and Sabre.

This rapid survey of the four possible referents for "R" technology is not exhaustive. The goal has been to illustrate how technology is creating an ever-more-personal, one-to-one, and one-to-many ecosystem.

Wrap-Up: Where Will "R" Technology Lead?

A review of the four stages successful sites move through ties up the basic message of this essay, namely relationships appear to create a suction that pulls users and keeps them stuck to the site. The " **network** effect" helped to make telephones, facsimiles, and electronic mail important to have. Broad use turned a curiosity into a necessity.

The behavior of the 180 million users worldwide suggests the **Internet** is simply a digital version of everyday human communications. Behind the glitzy Flash animations and the eye-straining typography of

portals, the **Internet** delivers a booster jet for talking in text or voice, in images, and **Web** applications that provide digital versions of catalogs, calendars, and airline ticket sellers.

After electronic mail, **search** and retrieval is the most used service on the public **Internet** . The difficulty of locating the right bit of information at the moment it is needed still plagues users.

The richness of innovation in **search** and retrieval is astounding. There are literally thousands of **Web** and online indexing systems with mind-boggling variations.

Access

Access is the starting point. Without access, there is no way for an individual to "get into" the space where the relationships exist. A person without an electronic mail address does not "exist" to some **Internet** users. A person unable to enter a particular site will have little or no knowledge about the interactions within that site. What information is available will be hearsay.

America Online has defended its proprietary Instant Messenger service fiercely. Access is limited to paying customers. Without value-added features like Instant Messenger, the appeal of for-fee services like America Online would erode. Similarly for-fee professional services place various restrictions upon access. High fees effectively prevent certain individuals and organizations from using LEXIS-NEXIS or Westlaw online information services. People without the funds to pay a hefty monthly fee are not allowed to use some of the financial information services available in the Thomsoninvest.com service.

Bonding

Savvy "R" technologists want to give those with access ways to form relationships with people who have similar interests and needs. Many **Internet** users take advantage of real-time chat rooms offered by the major portals. Millions of people post messages on various Usenet news groups, message boards, and special interest services. For just one example, look at the Beverage **Network** (http://www.bevnet.com), which features news, a bulletin board, and other information services for its users.

The rapid evolution of **Web** **search** services into full-service information portals creates rich assortments of information, communication, and interaction for users. Lycos' rapid rise from an also-ran to one of the top 10 most popular **Internet** sites has been fueled by its strong commitment to services that allow its users to buy sell, talk, trade, and auction in a global setting.

Other sites have discovered the power of real-time and asynchronous features as strong traffic builders. Talk City (http://www.talkcity.com), which began as a public forum for **Internet** users, has made substantial headway in **marketing** its conferencing and discussion services to corporations. It also functions as a public **Internet** site and an Application Service Provider for groups requiring limited-access discussion services. A similar surge in usage has accompanied the Webforia (http://www.webforia.com) initiative to provide vertical discussion services, e.g., a service to the information industry offered in conjunction with several well-known trade publications. Webforia may require users to register for some features. The company also offers low-cost **Web** software utilities for managing links and generating reports.

Mixing

With a wide range of clubs and messaging services, have sites reached the limits of simple communication? Several problems plague community, club, and chat functions. Many services are asynchronous. Users read and post messages at one time, returning at another time to look for new postings.

In order to make messaging more immediate, Yahoo!, Gooey, and other services provide visual cues when the author of a message or the person selling a product is online. Real-time interaction provides more solid evidence that the **Internet** is not a casual activity; personally or professionally. Real-time information becomes critical to people who use the **Internet** throughout the day without ever ending their connection.

eNow was one of the first companies to combine real-time messaging and real-time indexing of the information posted in eNow discussions. The well-known services of Deja News (http://www.deja.com) and Remarq (http://www.remarq.com) provide indexing of posts that often appear several

hours to several days earlier. (Deja News offers a product-oriented selection of newsgroup content and a broad index of general discussions; Remarq provides a keyword **search** of newsgroup postings. Both services use some filtering.) eNow provides near-real-time access to information so that a person monitoring an eNow session for competitive intelligence or time-sensitive information may have a jump on those using less timely indexing services.

Net Currents (http://www.net currents.com) provides a competitive intelligence service that includes features permitting information extraction from newsgroups, bulletin board postings, and other types of information not usually indexed by such spiders as Gulliver (Northern Light), Inktomi, or Lycos. Its remarkable combination of spiders and automatic **report** generation features monitors dynamic environments, gathering information about a company's financial position, postings and messages in various online groups, and news sources. On a schedule set by Net Currents' client, the company prepares a **report** about the company monitored. (A typical page from a May 2000 **report** appears in Figure 15 on page 48.)

In the "R" technology environment, the concept of "mix" is powerful. It connotes tools that allow individuals to mingle, exchange messages, comment, and create rich types of "homestead" information. ("Homestead" or "grassroots" information is created by individuals.

Unlike branded, third-party information from such sources as Reuters, the Associated Press, and The Wall Street Journal, among others, "homestead" information may or may not be accurate. It does, however, have value.) It also provides an ecosystem in which such advanced relationship technologies as **data mining** , spidering, and automatic **report** generation can thrive.

Nesting

Where do **Internet** customers spend their time? The goal of many **Web** sites is to build a loyal core of customers who habitually return to the site. The sites most adept at combining services and content for nesting are the most frequently visited.

The future of the **Internet** may be glimpsed in Yahoo!'s new service, FinanceVision. Before America Online announced its television service, Yahoo! had trumped the for-fee financial information services. A day trader can access a personalized **Web** page, near real time news and stock quotes, **Web** pages, a **search** box, and streaming video programs about companies and financial developments. The service is remarkable.

Emotional and Social Bandwidth

The **Internet** puts normal communications on steroids. It brings speed to what seems to be a basic human need -- communication.

Electronic mail is somewhat time-shifted. It lacks emotional and social bandwidth. The innovators who thrive in the **Internet** 's ecology are interested in software, systems, applications, and features that increase the warmth, individualization, and richness of **Web** interactions.

America Online's Instant Messenger brought ease-of-use and immediacy to text communication between "buddies." A number of companies created me-too versions of Instant Messenger. The success of America Online's application spread across customer service with robust commercial applications offered by Net Effects, Live Person, and eCRM (Electronic Customer Relationship Management) providers.

"R" technology boils down to using computer-assisted systems to bring human contact into what may seem to many a sterile experience. It represents a pragmatic response to adjusting a world that many people find far less comforting than the Norman Rockwell images of everyday life just 30 or 40 years ago.

As more nontechnical users embrace the **Internet** as a way to connect, the tools that make the user's life easier, more pleasant, or more meaningful will better integrate into that user's behavior.

Using technology to build, strengthen, or form relationships is a conceptual umbrella. The mechanisms of voice over the **Internet** or personalized greetings and recommendations are less important than the effect of these technologies on their users.

"R" is the suite of technologies to watch. In the next six to nine months virtually every major public site and most of the large intranet sites will place more and more emphasis on technologies that personalize, anticipate, provide enriched communications services, and, most

importantly, build "me.coms" for people with affinity.

Stephen E. Arnold of Arnold Information Technologies (AIT) (http://www.arnoldit.com) and Michael Colson of Talavara LLC (http://www.talavara.com) both provide consulting services in "R" technologies and **Web** commerce.

PUBLISHER NAME: Information Today, Inc.
COMPANY NAMES: *Yahoo! Inc.
EVENT NAMES: 010 (Forecasts, trends, outlooks); 600 (**Market** information - general)
GEOGRAPHIC NAMES: *1USA (United States)
PRODUCT NAMES: 7372422 (DBMS Utilities); 7375000 (**Database** Providers)
INDUSTRY NAMES: LIB (Library and Information Science)
SIC CODES: 7372 (Prepackaged software); 7375 (Information retrieval services)
NAICS CODES: 51121 (Software Publishers); 514191 (On-Line Information Services)
TICKER SYMBOLS: YHOO
SPECIAL FEATURES: LOB; COMPANY
ADVERTISING CODES: 57 New Products/Services
?

8/9/1
DIALOG(R)File 15:ABI/Inform(R)

01730804 03-81794
Mine over matter
Baker, Sunny; Baker, Kim
Journal of Business Strategy v19n4 PP: 22-26 Jul/Aug 1998 CODEN: JBSTDK
ISSN: 0275-6668 JRNL CODE: JST
DOC TYPE: Journal article LANGUAGE: English LENGTH: 5 Pages
WORD COUNT: 3030

ABSTRACT: Just 5 years ago, only 50 researchers received Gregory Piatetsky-Shapiro's monthly electronic newsletter, Knowledge Discovery Nuggets. Today, the newsletter has 4,000 readers and is published 2 or 3 times a month. **Data mining** - the use of statistical methods and new **search** software to uncover useful patterns inside databases - continues to attract more attention in the **business** and scientific communities. There are massive amounts of data in corporate coffers that could be used to reinvent marketing strategies, and **data mining** is one way to find information that matters.

TEXT: Headnote:

There are obscene amounts of data In corporate coffers that could be used to reinvent marketing strategies. **Data mining** Is one way to find the information that counts. 1

PITY THE **MARKET** RESEARCHERS FROM PREVIOUS decades. Blessed with ungainly **database** design, slow throughput, and a host of compatibility problems, gathering even simple customer demographics required days-long model-building exercises to launch only the crudest **search** for useful information. Assembling a data sieve required help from the MIS gurus who, with a Dilbert and Wally level of marketing savvy, failed to understand the purpose of pulling facts from the data archives. Underfunding the **data mining** effort was common because what senior managers assumed would take minutes really took weeks to plan, implement, test, analyze, and summarize. 2

The resulting multi-page reports were of more use to paper recyclers than to information-hungry decision makers. In those dark days, **data mining** had much in common with coal mining. It was dark, awkward, gritty work with perilous hazards that could endanger the life of the **business** . 3

Just five years ago, only 50 researchers took part in a knowledge discovery and **data mining** conference workshop and also received Gregory Piatetsky-Shapiro's monthly electronic newsletter, Knowledge Discovery Nuggets. Today, however, the newsletter has 4,000 readers and is published two to three times per month. And **data mining** -the use of statistical methods and new **search** software to uncover useful patterns inside databases-continues to attract more and more attention in the **business** and scientific communities. 4

According to Piatetsky-Shapiro, still the editor of Knowledge Discovery Nuggets and also a director of applied research at Knowledge Stream Partners, a **data mining** consulting company based in Chicago, "We are clearly moving into the next generation of **data mining** systems, toward more and more embedded solutions." Over the last two to three years, there has been more emphasis by developers on integration, visualization, and data access tools, and the number and sophistication of these tools is increasingly rapidly. And in a 1997 **report** , Stamford, Conn.based Gartner Group stated: " **Data mining** and artificial intelligence are at the top of the five key technology areas that will clearly have a major impact across a wide range of industries within the next three to five years." 5

Companies now use computers to capture details of **business** transactions such as banking and credit card records, retail sales, manufacturing warranty, telecommunications, and myriad other transactions. The data from these transactional systems contain thumbprints of the key trends that 6

affect various aspects of each **business** -including information about products that sell together, sources of profits, factors that affect manufacturing quality, and other mission-critical relationships. The predictive relationships and trends buried in the data are the gold revealed through **data mining** .

Fueled by the increasingly rich data sets maintained and often underused by corporations, a new generation of **data mining** tools from companies like AbTech, **Data Mining** Technologies, Future Analytics, Information Discovery, NeoVista, Ultragem, and others, are moving data analysis out of the MIS department and closer to the boardroom. While it's true that a complicated **search** still demands the services of a data analyst or two, new tools enable managers in many companies to conduct much of the work by themselves with only a modicum of training to make wise and informed decisions. 7

Today, the newer and easier-to-master **data mining** tools can become a senior manager's Rosetta Stone for translating even hundreds of terabytes of data into useful models (a terabyte is a 1,000 gigabytes, and even one is a lot of information to sort through). A dig in the right place with the right tools translates flat **database** gibberish into gems of insight about the organization and its customers. 8

Tomorrow Is Today

Because large databases often provide too much of a good thing, approaches based on traditional query languages and OLAP (on line analytical processing) usually encounter a problem known as "The Maze of a Million Graphs" where a user can build a million charts and still not see the forest for the trees because there is so much data. **Data mining** , on the other hand, draws its power from software tools that **search** through the data intelligently to look for patterns and relationships. The best software, on its own initiative and using embedded algorithms, discovers key patterns in the data warehouse by itself. 9

Data mining tools extend decision **support** capabilities that allow managers to query information in databases and turn the results into reports. The difference is that **data mining** tools look at vast amounts of data, often from multiple sources, and find patterns and relationships in the data that would otherwise remain obscured from ordinary **database** queries. 10

Data mining goes hand-in-hand with the new data warehousing tools that are necessary to organize the historical information gathered from large-scale client/server-based applications. For example, **data mining** tools offer a way to identify, extract, and analyze the valuable information contained in the massive databases that are often a byproduct of on-line transaction processing (OLTP) systems. The data from the on-line transactions, after being appropriately "cleansed" and organized in a relational **database** format (commonly referred to as a data warehouse) can then be mined to discover patterns in customer activity. These patterns provide **business** analysts with insights into customer behaviors that would otherwise be indiscernible. The insights lead to proactive, optimized **business** initiatives to reach, serve, and target customers. 11

What Can You Find in a **Data Mining** Expedition?

Storing embarrassing amounts of data is something that computers excel at. Getting it back is a different matter. Without intelligent software intervention, snagging the right data is impossible. And only within the past few years have easy-to-master **search** tools become affordable and powerful enough to get information onto ordinary desktop computers. 12

The best **data mining** tools generally solve one or more of the four basic problems in turning warehoused data into predictive information: 13

Classification of data into meaningful entities.
Classification results in defining the data elements that exhibit similar behaviors or relationships to other data.

Modeling explicit dependencies between variables, such as the dependent relationships between income and nonessential purchases, to cite a simple example.

Clustering data into a finite set of categories (clusters of data variables) that describe the data relationships in terms of predictable actions, performance, or behaviors. 14

Detecting deviations in key data from previous or expected values, which allows data miners to use deviations from the past to predict changes in future trends, output, or behaviors. 15

Data mining tools that solve these data problems produce predictive knowledge from data warehouses. For example, in customer prospecting and segmentation applications in industries such as banking, credit cards, and insurance, **data mining** helps break the **market** into meaningful clusters and segments. This helps identify customer groups that are not only more likely to respond to offers, but also will provide better profits in terms of higher volumes of product usage. And, by analyzing results of direct mail campaigns, **data mining** helps identify higher-value responses (in terms of profit per mailing) as well as higher response rates. This information helps marketing managers and **business** executives focus their promotional activities and new marketing campaigns on the opportunities with the biggest payoff. 16

In customer relationship management, **data mining** finds patterns of product usage and consumer behavior. It helps managers understand what causes customer attrition and what will improve customer retention. It uncovers crossproduct utilization behavior, improving the management of channels such as bank branches, automated teller machines, and service outlets. These patterns can reshape the thinking of a customer loyalty department. And, because **data mining** can combine trend analysis with wallet-share and affinity results, companies can design better marketing programs that can be implemented based on a customer's life cycle model. 17

In the retail sector, **data mining** helps identify how products sell together in specific stores or regions through affinity and **market** basket analysis. **Data mining** also helps companies understand profit patterns per linear foot of shelf space, based on store layouts and product combination offerings, and determines when promotions work best and which item layout combinations are the most profitable. **Data mining** also helps identify products that are traffic builders and those that are profit makers. Retailers can assess the impact of advertising and in-store promotions and more accurately analyze inventory control issues. 18

Data mining efforts in companies specializing in consumer packaged goods are the mirror image of retail data analyses. These **data mining** expeditions analyze one manufacturer's products in many store chains, instead of several manufacturers' products in one chain. Here, **marketshare** analysis by store-chain and demographics, and understanding where and when promotions work and when they don't helps focus advertising dollars. And, by analyzing historical performance from the chain level down to the region and store levels, companies can identify the factors influencing the effectiveness of advertising. 19

Manufacturing quality programs can use **data mining** to help manage interactions among a large number of variables. **Data mining** tools automatically identify the unusual data densities that are the tell-tale signs of process variations within manufacturing and assembly operations. For example, companies can automatically identify the combination of plant, build-date, and product models that have a higher malfunction rate, allowing quality engineers to quickly focus on the source of problems. **Data mining** can also help improve product quality by identifying the factors that give rise to problems. 20

In the telecommunication industry, **data mining** identifies the patterns of change in the **market** , allowing the marketing department to better 21

focus on customers who demonstrate a high acceptance of services and longer usage. **Data mining** also helps with capacity planning by providing an understanding of the underlying patterns and structures of service usage by customer groups. This insight allows capacity planners to optimize the investments in network facilities to better serve customers, while avoiding costly overexpansions.

What Do You Need?

All the **data mining** applications mentioned already, and more, including programs for direct marketing, fraud detection, and stock **market** prediction, depend on reliable software programs that integrate with the corporate data standards. The program or suite of programs you choose should be fast, adaptable to your specific **business** needs, easy to learn, simple to use, and able to present results in terms that non-statisticians can understand. You should also buy the software from a company that has a good reputation for **support** , training, and consulting services, because even the best **data mining** software needs explanation. The vendor you choose should be willing to show you the ropes for going down into the data mines. 22

Beyond the right software, there are four additional components required for successful **data mining** :

1. Quality data organized into an accessible and extendible data warehouse. The data must be of the right age and richness (depth) for the task. Older data is often useful for studying trends, so age is not necessarily a criteria that renders it useless, even though the same data might be almost worthless when assembling a simple mailing list. 23

2. A well-executed model for extracting the data. While most mining operations aren't that difficult to assemble, a mistake is a two-edged sword. Snag the wrong data and the results may appear credible while completely missing the most valuable relationships among the records. Grab too much data in a model and the results become too cumbersome to produce a useful study.

3. A way to verify models and results. Testing preliminary results is a must for ensuring the project doesn't run off the rails for something a simple as a syntax error or as complicated as flawed logic. Analysis and testing of the completed results is also a must. A good **data mining** software company will provide access to training and consulting services to assure that your searches for gold in your data produce the pay dirt you're looking for.

4. An accessible reporting function. The output from **data mining** software should make a persuasive case for the results of the study. The best reports incorporate statistical data for analysts, coupled with a simple but convincing summary of data implications for less technical management.

Companies moving into **data mining** must first inventory the databases they own or have access to. Understanding what's already on tap spurs ideas on integrating the information for problem solving and customer management. Many companies have more data than they realize, often spread among departments as well as the MIS facilities. An inventory and merger of these data sources may result in eliminating previously duplicated databases and correcting a surprising number of mistakes. 24

There's no need to rely on internal data alone, however. Today, companies can buy, rent, or even borrow data from so many sources it's mind boggling. A major pizza chain looking to open new outlets might rent data from one of the many demographic data providers. A large corporation might combine its own data with commercially available data to analyze **business** customer needs and assemble a model to study historical trends and, from that study, decide which of several product prototypes to fund. 25

Industry surveys indicate that over 80% of Fortune 500 companies believe 26

that **data mining** will be a critical factor for **business** success by the year 2000. Most of these companies now collect and refine massive quantities of data in data warehouses.

These companies realize that to succeed in a fast-paced world, **business** users need to be able to get information on demand. And, they need to be pleasantly surprised by unexpected, but useful, information. There is never enough time to think of all the important questions. You need a computer with **data mining** software to help find answers to the questions you don't have time to ask. Such **data mining** can provide the winning edge in **business** by exploring the databases on "automatic" and bringing back invaluable nuggets of information. 27

If you're not already involved in **data mining** , it's time to get out your computer-based pick ax and start digging for the gems in your data. And if you don't, you can be that someone else will. 28
Sidebar:

Data Mining Software Companies

The following is a short list of companies that provide **data mining** software solutions For more companies and information on **data mining** , use the keyword **data mining** ! with the Yahoo! **search** engine on the WorldWide 29

AbTech Corporation 1575 State Farm Boulevard Charlottesville, Va. 22911-8411 limB: 804-977-0686 FaX: 804-977-9615 Email: sales@abtech.com 30

AbTech's ModelQuest sup (R) mining tools are based on a number of advanced machine learning technologies and AbTech's proprietary StatNet Expert approach. StatNet Expert technology is a **data mining** approach that has evolved from more than 30 years of research in developing predictive data modeling solutions to real-world problems. It combines the neural net, regression, and expert system methods and automatically captures the complex, non-linear relationships in data. 31

Data Mining Technologies, Inc. 1500 Hempstead Turnpike East Meadow, N.Y., 11554 Phone (516) 542-8900 Fax: (516) 794-4672 32
Email: info(at)data-mine.com

Data Mining Technologies is a new company that produces Nuggets, which automatically sifts through data and uncovers hidden facts and relationships. Nuggets can reveal which indicators most affect your **business** , and help predict future results. It works across industries and is effective with a wide range of applications. According to the company, it can help identify the best **market** segments, predict product success, reduce fraud, assist with credit scoring, forecast equipment failures and maintenance needs, improve manufacturing quality control/fault analysis, and assist with medical studies and scientific research. 33

Future Analytics, Inc. 7 West Washington Street P.O. Box 1455 Middleburg, Va. 20118-1445 Phone: (540) 687-3692 Fax: (540) 687-3654 34

Email: info@futureanalytics.com

Web Site: http://www. futureanalytics.com/ Future Analytics, Inc. provides cutting-edge analytical services to **business** users. Its core technologies consist of **data mining** , statistical analysis, data warehousing, and Web enablement. The company uses technologies such as neural networks, genetic algorithms, and decision trees to solve complex analytical problems. The company's professionals help companies design research studies, create data collection forms, and carry out statistical analyses. This service can be combined with **data mining** and data warehousing components. The company 35

Sidebar:

provides an integrated service, from warehouse assessment ants conStruction to warehouse exploitation.

Information Discovery, Inc. 703-B Pier Avenue Suite 169 Hermosa Beach, Calif. 90254 Phone: (310) 937-3600 Fax: (310) 937-0967 Web.. http://www.datamining.com 36

Information Discovery, Inc is a leading provider of largescale **data mining** -oriented decision **support** software and solutions. Its products serve all the major decision **support** needs with pattern discovery and **data mining** software, strategic consulting, and warehouse architecture design. The company offers a variety of customized **business** solutions to augment the **Data Mining** Suite and the Knowledge Access Suite products for enterprise-wide, large-scale decision **support** . Using these products, companies can directly mine large multi-table SQL databases with no need for sampling or extracting files. And the Knowledge Access Suite includes a gateway to knowledge that has been pre-distilled from data and stored in a pattern-base. **Business** users need not perform data analysis, but simply query explainable knowledge on the intranet that has been automatically pre-mined. 37

NeoVista Software, Inc. 10710 North Tantau Avenue Cupertino, Calif. 95014 Phone: (408) 777-2929 Fax: (408) 777-2930 EIa:webmaster@neovista.com 38

According to company representatives, NeoVista Decision Series' knowledge discovery methodology can expose the patterns that are critical to **business** success to provide an advantage the competition may never discover. NeoVista's Decision Series suite of **data mining** tools enables companies to detect the patterns and trends in corporate data that lead directly to predictions of customer behavior, a targeted marketing focus, improved operational effectiveness, and optimal return on investment. NeoVista's Decision Series products integrate with open relational databases and standard SMP parallel hardware platforms. 39

Ultragem **Data Mining** 450 Wildberry Drive Boulder Creek, Calif. 95006 Phone: (408) 338 3302 Fax: (408) 338 7503 EmI:mail@ultragem.com

Ultragem uses the company's proprietary high-performance genetic **data mining** technology. Genetic **data mining** is the automatic extraction of prediction and classification rules from databases using advanced algorithms supported by the Ultragem software. For example, **data mining** can discover rules that will accurately predict the probability that a borrower will default on a loan. Genetic **data mining** can discover prediction rules in data which no human mind could ever have found, and which were beyond the reach of earlier technologies. 40

Author Affiliation:

Sunny Baker, Ph.D., is associate professor and director of technology for East Tennessee State University. Kim Baker, a marketing consultant for high-technology companies, is the author and co-author of more than 20 books.

GEOGRAPHIC NAMES: US

DESCRIPTORS: **Data mining** ; **Market** research; Statistical methods; Data bases
CLASSIFICATION CODES: 9190 (CN=United States); 5240 (CN=Software & systems) ; 7100 (CN=Market research)
?

4/9/8
DIALOG(R)File 16:Gale Group PROMT(R)

06938168 Supplier Number: 58545238 (THIS IS THE FULLTEXT)
The Answer Machine.(information services management)(Industry Trend or Event)
Feldman, Susan
Searcher: The Magazine for Database Professionals, v8, n1, p58
Jan, 2000
ISSN: 1070-4795
Language: English Record Type: Fulltext
Document Type: Magazine/Journal; Professional
Word Count: 12581
TEXT:
When we **search** for information, we want answers, not documents. Current retrieval systems find documents that may or may not contain the answers to the questions users ask. In the next 5 years, perhaps sooner, information systems as we know them will change dramatically. These systems will find real answers, moving from the static to the dynamic, using machine learning techniques to adapt to new information and to new interests. Finally, these systems will learn to interact with the user, delivering information in visual, easy-to-understand packages that can be manipulated and used collaboratively. 1

Datasearch

This information revolution is fueled by increased demand, by improvements in computer technology, and by our growing comprehension of how people seek and use information. As non-information professionals have become the dominant information consumers, they have begun to demand systems that can locate and manipulate information without arcane command languages and other traditional priestly rites. Unlike information intermediaries, whose main function is to **search** , knowledge workers use searching as a means to an end. This increasingly sophisticated group of information end users needs to find the right information quickly, analyze it, combine it into reports, summarize it for upper management, or use it to make decisions. They need a suite of integrated, intelligent information tools to make sense of today's ceaseless information bombardment. 2

Faster, bigger, cheaper desktop computers have the capacity to run newly developed information handling tools. News information systems will be built upon a foundation of linguistic analysis of language and meaning. To this, we add our growing understanding of cognitive processes. Research into how people think, combined with observations of how they interact with computer systems, is spawning the new discipline of human-computer interaction. New systems will draw heavily upon this field, as well as on cognitive psychology, graphic design, linguistics, computer science, and library science, each system with its own unique perspective on how to organize, find, and use information effectively. 3

The growth of corporate intranets adds to the demand. Companies are willing to invest in high-end, carefully crafted systems. **Business** cycles are growing shorter, while pressured employees spend too much time trying to handle too much information. Knowledge walks out the door as employees leave for new jobs in other companies. Intranets attempt to preserve this information and make it available to the entire company and will become an interactive venue for working with colleagues and with information in one smooth process. 4

Today's document retrieval systems lump all information needs into a single process. New information tools will separate these different needs into categories and provide specific tools for each kind of need.

Here are some of these **search** types: 5

* Broad subject searches -- fishing expeditions about a topic unfamiliar to the searcher. Appropriate terminology is hard to determine at first.

* Narrow, well-defined subject searches on a familiar topic with known terms.

* Comparative, information-seeking -- which company is the biggest, has revenues of more than $X, or more than 100 employees?

* Known-item searching for a specific title, author, or publication.

* Continuous monitoring of a subject.

* Pattern matching for emerging trends: foraging for matches to a description of an event or a profile of a competitor or other entity.
* Fact or statistic location -- who, what, where, when, how?
* Chronological reconstruction of events or actions

The **Search** Process

How do we **search** for and use information? Do end users differ from information intermediaries, and if so, why? Can we differentiate types of searches and develop specialized tools to improve our finding and use of information? These questions and more must be answered as we Set about designing the next generation of information systems. 6

Searching isn't linear. We know that people engage in an iterative, or circular process when they seek information (see Figure 1 on page 61). 7

After testing the **search** behaviors of both end users and information professionals for the last 5 years, I believe in the inherent differences between how both groups **search** . This is not surprising, but it has little to do with the skill or training of either the information professional or the end user.' Rather, these groups differ in their fundamental motivation for searching. End users know why they are searching, even if they don't articulate their information needs well. Success is defined by an answer to their information needs. They will know it when they see' it. Therefore, they will more likely enter a very broad query, and then browse. In fact, given a choice, they will enter the **search** cycle by browsing first and then refining their browsing with a query. This explains the popularity of directory sites like Yahoo! 8

In contrast, the intermediary has only the end user's question to match. Success is defined by the best possible match. Therefore, intermediaries focus on precision. Their queries tend to be much narrower and they will **search** before they browse. A broad query to the information professional is unprofessional, sloppy. When we criticize end, users' for their lack of searching artistry, we are' often mistaken. They need to browse and browse broadly (see Figure 2 on page 67). 9

Most of today's document retrieval systems match queries to documents. 10

These systems address the middle of the information-seeking process, en closed in the dotted lines. While we may complain about the results, in fact, the systems do a pretty good fob of matching the actual query received. However, the systems ignore the two outer ends of the process, offering no help at all in translating information nets into questions and: then into acceptable queries. The systems do little to help the user understand and analyze what the system returns. So, while the user actually receives some good matches to his query, the query rarely reflects the information need behind it. 11

Yet, if the information need is not represented accurately, then the results returned will at best intersect that need spottily. Today's information sys terms require the searcher to extract terms that have the best chance of representing the question, while at the same time, eliminating extraneous or unrelated documents. We usually resolve this dilemma by using lists of nouns or phrases that represent the concepts in the question. In the process of formulating a query, we eliminate the actual meaning of the question because we strip away the context. 12

Look at the list of questions in the "Stinkers" sidebar at left. A real Answer Machine could answer these questions, and more. It should:

* Help the user formulate a query.
* Find answers, not just documents.
* Anticipate user needs.
* Retrieve data in any format, from multiple sources and suppliers. and merge the data into a single, de-duped retrieved set. 13
* Provide analysis and reporting tools to manipulate retrieved data.
* Display answers in easy-to-digest visual formats.
* Find patterns within data to **support** decision-making.

This is not as far-fetched as it sounds. Most of the, technologies that the Answer Machine requires are already in development. Answer machines will become the technical underpinnings of knowledge management systems, providing single, organized, easy access to all the information in an. organization including the following:

* Internal documents and databases in a variety of formats
* Committee reports
* External sources

* Directories of people and skills
* Tools to manipulate information and extract new knowledge

The trick will be to select the appropriate tools and then to present them as a seamless system. All the technologies discussed in this article should be thought of as pieces of a whole: a new model for a information system that brings together all of the resources of an organization in any format. 14

If you approach your information system as a whole, then you will implement each new technology as a brick within an entire edifice. You could implement each technology separately, but ultimately, integration, of these technologies will create a knowledge management system arid even a decision **support** system. Without this vision, you may end up with so many oddly sized bricks that you will have to start again from scratch. 15

The system you build should adapt to user needs and integrate information in any format. It must reveal patterns and trends in information, because patterns, and trends are usually more significant than discrete.facts or nuggets. And above all, it must deliver answers to questions. 16

The Foundation

Any retrieval system must distinguish between one document and another. The system relies on indicators that determine what a document is about. It also tries to differentiate between documents "mostly about" a topic and those merely "somewhat about" a topic. Unique terms or phrases often serve as good discriminat6rs. However, unique terms are hard to find in some areas, such as **business** , which use very common words to mean something quite precise. The sample queries "Stinkers" offer good examples of this problem. 17

To best determine a document's meaning, ask a subject expert. Indexers do this f6r a living: However, while experts may agree On broad subject areas, they may differ On which terms to assign to a specific document. So the studies done on indexer consistency found that indexers assign the shine term to the same document only 50 percent of the time. Indexers do classify documents in the correct general subject area, even if they don't assign precisely the same term. They don't put financial institutions under environmental science. 18

Why, then, don't we stick to human classifiers to determine what a document is about? There are several reasons. First, that 50 percent consistency rate is quite troubling if searchers use thesauri to aid in query formulation. 19

Assigning the wrong term can eliminate a highly relevant document from a retrieved set. Second, human indexing is slow; it adds weeks, even months, to the time it takes to make something available online With real-time publishing becoming an accepted practice, we need other reliable means of distinguishing the relevant from the irrelevant. Third, the sheer volume of information is too great to try to classify it all manually. 20

Given that we must find an automatic means to select the best documents for a query, how can we teach a computer to recognize a good match? 21

Statistics and Probability

For all that searchers talk about words, terms, commands, and other linguistic phenomena, computers really understand only numbers. Every ASCII character, every letter in the alphabet, must be translated into a sequence of ones and zeroes before a computer can crunch it. Boolean commands work quickly because they are mathematically based. One of the ironies of online searching is that, its practitioners consider themselves to be "word" rather than "math" people. Yet, they handle Boolean logic with aplomb. 22

The genius of people like. Gerard Salton lay in their recognition that text contains predictable patterns. These patterns can be described mathematically, so that computers can detect them and then perform statistical and mathematical operations on them. For instance, it seems, obvious that the more a document is "bout" a subject, the more times words dealing with that subject will appear in the text. Conversely, these terms should not appear very frequently in documents not "about" that subject. This is the rudimentary idea behind relevance ranking in retrieval, systems. 23

Clusters of certain terms are even better indicators, that a document is about a particular subject. The appearance of co-occurring terms will determine more precisely when a topic is central, to a document. None of 24

this requires that we understand the meaning of the words, merely the patterns the words display in the text.

Needless to say, we could embellish this principle by saying that words in the title are more important than words in the body of the document. We could add that the closer together subject-relevant words appear, the more likely the document is about what we are searching for. Or, if the words appear in a lead paragraph, they are more important indicators of the subject than if they appear in paragraph five. This is what skilled searchers do in crafting a **search** . It is not magic.

If we can describe these patterns, we can program a computer to find them. The first mathematical operation that **search** engines do is to count, something that computers do very well and very fast. Computers count the number of times a term or terms appear in a document, then assign a weight, or number, that represents this count to distinguish one document from another. This weight calculation, usually takes into account how rare the term is in the whole **database** -- how many times it appears in every document in the collection. Rare terms are often good discriminators and receive a higher weight.

Search engines may also truncate terms to include plural and singular forms. Extra weight often attaches to terms appearing in the title or lead paragraph, as to documents which contain several query terms in the same sentence or in the same paragraph. Most **search** engines also, "normalize" results to take into account variations in the length of documents, since longer documents will probably contain more occurrences of a term. When a **search** system matches your query terms to documents, it adds up the weights for each query term that appears in a document and assigns a score for that document. Then it compares all, the scores and presents the highest first. This is relevance ranking in a nutshell.

Statistics and patterns enter into advanced retrieval systems in a number of other contexts. For instance, in order to determine whether a document matches a query, the system must calculate the similarity of the document to the query. The human mind does this without trotting out an algorithm. Computers must translate both query and document into some sort of representation. About this task, experts have written whole books.

One approach is to translate both query and document into a "vector" - a line which goes off at a specific angle from the center of an imaginary space. Think of this space as having a signpost at the center, with each individual sign pointing in a slightly different direction. The words in the document all point to specific directions in this imaginary landscape. Documents containing similar words will point in the same general direction; the more similar those document terms, the closer their angles will be to each other. We can measure these angles to give us a degree of similarity. This "vector space model" can help calculate relevance ranking, but it can also determine clusters or clumps of similar documents. This is the basis for most of the star maps or imaginary. Landscape visualizations used to display the contents of a **database** , or a retrieved set of documents.

These, statistical techniques work surprisingly well in the majority of cases. But these techniques do not work well for every query. That is the nature of statistical methods. When we hit an exception to the rule, the errors can be, glaring, unlike human errors. For instance, when a query contains both a very important concept expressed in an extremely common term and a very minor concept expressed in a rare term, then the rare term may skew the relevance ranking, since it has a higher, weight than the common term.

Remember also that statistical systems do not "understand" a query, but operate on the numbers. Many meanings for the same word elude this kind of technology Financial institutions may be classified as environmental science, if the word is "bank." However, since bank will not appear in combination with other environmental terms, if a query is more than one word long, a statistical system Would rank such a false drop low. Hence, **search** engines look very stupid by making errors that any human with half a brain would never make. This could explain why **search** engines have such a bad reputation among most professional searchers; their errors are unreasonable. That is because the meaning of the terms being retrieved is not part of the equation for statistical processing.

Natural-Language Processing

In order to build a state-of-the-art information system, one must

extract as much meaning as possible from each document. A list of words, or even words and phrases, is not enough. Context and meaning must be preserved. Only a system able to distinguish meaning can return articles about terrorists instead of rugby matches when asked for attacks, skirmishes, and battles in Rwanda. A meaning-based system will also know to return predictions about future, not past, production of widgets in Zambia in Question 9 of our "Stinkers" list. 32

To create an advanced information system, first one must build a knowledge base. This base will contain all the documents in the system and their words, but also added information to resolve meaning and dissolve ambiguities. A good natural-language-based system provides the foundation for this system, because it parses sentences thoroughly, extracts meaning from context, and is smart enough to realize that if the year is 1999, Hilary Rodham Clinton and the first lady are the same person. A document-processing tool is required that can extract and **store** many layers of meaning, as well as automatically categorizing documents and identifying all variants of proper names. Each unit of meaning may also carry a time stamp relating to the content, not to the date on which someone added the document to the **database** . With relevant dates in place, later tools can extract automatically chronologies of events. Chronological information also enables the system to distinguish between first ladies Barbara Bush and Hilary C linton, depending on the time and context of the question. The information in the knowledge base should also be retrievable as separate units, such as a single sentence or paragraph, if we want it to supply direct answers to questions. 33

I stress this knowledge base building step because most organizations will not willingly invest the money, time, and effort needed to design a knowledge base more than once within a few years. Any future advanced information tool will operate on the contents of this knowledge base. Therefore, extracting as much knowledge as possible should increase the flexibility in the future to adopt new technologies, as they arrive We can't know now) what tools in what formats; research and the **market** will deliver in the next 5-10 years Compatibility will always be an issue However, raw knowledge does not change. The more handles that you create to grab a Piece of information, the more chance that you can retrieve it when needed. This is the same principle that advises digitizing at a high resolution when scanning collections: Build the foundation wisely and richly, because you'll never be able to start again from scratch. 34

As we build advanced information systems, we will require that the systems understand text as we do. Natural-language (NLP)-processed-based systems are the only ones to answer this description at present. While NLP systems match terms, Las both Boolean and statistical **search** engines do, the systems also extract meaning from syntax, built-in lexicons, context, and even the structure of the text itself. This is what humans do to figure out what a document means. 35

Many people feel that statistical and NLP systems won't work as well on bibliographic databases because their forte is full-text searching. True, these systems are not designed to work well on document records that do not contain substantial text. Therefore, it is said that bibliographic records such as those appearing in typical library are not good candidates for these advanced retrieval systems. However, I have found these systems as effective as Boolean systems in searching through bibliographic databases, because most can default to a relaxed Boolean query if necessary As an added benefit the ability of the systems to relax the strictures of a query means that occasional typographic errors will be ignored in relevant records that a Boolean system would eliminate from the results. 36

Intelligent Agents

Imagine an information system that learned what you sought and began to anticipate what you would like to see. While this may sound like star Wars, in fact, this capability exists in embryonic form today. Interactions with today's systems are fixed in time. The searcher must change a query in order to find documents not already retrieved and to add new indexing terms manually. We need systems that adapt to both the changing interests of the user and to changes in the terms used to describe each topic. Machine learning techniques can make an information system dynamic. 37

For instance, suppose 3 years ago you set up an alert for anything on "information retrieval." If you didn't; change your Alert profile, you would miss all the articles on **data mining** , knowledge management, or 38

automatic summarization. An intelligent agent system could detect the rise of these new terms. The system would find clues in the appearance of **data mining** as a co-occurring term with information retrieval. Or, the agent system might note that you were reading articles on **data mining** and ask if you wanted to add that term to your profile. It might be programmed to follow new **Internet** links from sites that interested you; or, it could run an updated query periodically on all the **Web search** engines and then follow those links This is of immense importance in a world in which, in 1997, a Reuters survey found that most professionals spent more time seeking information than using it.

Intelligent agents are software programs that use machine learning. Agents do not have innate intelligence. Although agents can operate in situations that have underlying patterns or rules of some sort, agents cannot work in complete chaos or with random input. The patterns or rules that they rely on may be described by humans or developed by the agent-based system itself. An agent system develops rules from sets of representative data and queries -- a training set. During the training period, system agents "learn" the best matches by trying out various matches and receiving corrections from human input. Eventually, agents build a pattern for what constitutes a "good match." 39

Agent systems are autonomous -- in other words, they can initiate actions within a carefully defined set of rules. They are also adaptable, able to communicate with other agents. and with the user. Agents may be mobile, traveling along the **Internet** or other networks in order to carry out various tasks, such as finding or delivering information, ordering books, or monitoring events. Most importantly, agents can alter their behavior to fit a new situation. They learn and change. 40

Some agent systems exist today. See the Botspot (http://www.botspot.com) for an extensive list and description of such systems. The agents in the Microsoft Office suite are only a beginning. They are not adaptable and they follow set rules. These agents offer hints, take and sometimes answer questions about functions of the software, and are mildly amusing. Eventually, we can expect agent systems to adapt to our preferences for formats or other repetitive actions we take -- like opening applications in certain orders or checking e-mail at a certain time of day -- and will perform these tasks automatically. 41

Eventually, agents will play a big part in the decision **support** systems now in development. These systems will use a knowledge base to find and compare previous situations that might apply to current problems, offering alternative solutions and perhaps creating scenarios for each alternative. 42

These three disciplines -- statistics, natural language understanding, and intelligent agents -- form the foundation for understanding and using the information tools of the future. While it will be possible to use these tools and never understand their inner workings, those who delve below the surface rules will use them most effectively. Apparent anomalies and mistakes will be come less puzzling as well. 43

NLP-Based Technologies

By examining meaning instead of just matching strings of words, NLP

Systems can solve many retrieval problems intelligently. These include identifying concepts, even if different terms are used to describe the same idea. NLP systems should identify the names of people, places, or things in any form. The systems could also encompass speech processing, summarizing documents, and even groups of documents, and automatically indexing and classifying documents. Each of these aspects represents a distinct area of research with tools in development or, in some cases, already on the **market** . 44

Concept Extraction and Mapping

Concept mapping is the key to many new technologies on the horizon. Language provides rich alternatives in how an idea is expressed. Not only are there direct synonyms, but metaphors, similes, and other literary devices. These devices delight the reader, but puzzle the computer. We need systems that can use all those levels of language to interpret meaning correctly and to relate similar expressions of an idea to the same concept. 45

Concept mapping enables us to:

* **Search** across disciplines using different vocabularies to express the same idea.
* **Search** across languages for the same subject. 46

* Identify and retrieve all variants of a name or place, no matter how a question is phrased.

* Index materials automatically.

Concept and vocabulary mapping are like creating a controlled vocabulary. In a controlled vocabulary, all synonyms are identified and one is chosen as the "official" term. Other terms cross reference to that official term. Concept mapping works in a similar manner, except that the concept does not need to be a single chosen term. Instead, all synonyms form a cluster of terms that represent the idea. Since the idea is represented abstractly, it can cover not only words in one language, but in any other language and well beyond the conceptual grasp of multilingual human dictionaries or thesauri. 47

Vocabulary mapping, a form of this technique, enables a searcher using MESH terms in MEDLINE to **search** intelligently in CINAHL, another medical **database** with a different thesaurus in control. Thus, the idea of "tree" has multiple terms mapped to it, as shown below in Figure 3. 48

This is a technology already in place with varying degrees of sophistication. It is used in the following areas. 49

Machine-Aided and Automatic Indexing

Machine-aided or automatic indexing (MAI) finds major concepts in texts, maps them to an internal thesaurus or controlled vocabulary, and applies indexing terms automatically. It may also extract important names, disambiguate words, and identify new terminology for indexers to add to the system. MAI offers candidate terms to indexers for their approval. Automatic indexing applies these terms with no human intervention. 50

Machine-aided indexing has been around a long time. Most such systems are rule-based and assign terms based on rules such as "use 'automobile' as an indexing term whenever a document is about 'car'" just as professional human indexers do. Data Harmony/Access Innovations is well known for its rule-based machine-aided indexing systems. Northern Light uses rules developed by human indexers to automatically assign broad terms to all documents for its customer folders. Autonomy uses machine learning to automatically categorize materials, and Semio creates taxonomies or hierarchies automatically. Systems such as DR-LINK, developed by Dr. Elizabeth Liddy at Syracuse University, assign subject codes in order to disambiguate words. Some MAI systems work with up to 80 percent accuracy, which compares favorably with manual indexing. 51

Some experimental approaches use probability and statistics to categorize materials. Muscat, now owned by Dialog, is a good example of this approach. Others are experimenting with neural networks for automatic classification. 52

MAI systems can also extract important names from the text or "disambiguate" terms. Consider the term "bank." It may be a place to **store** money, the side of a river, a turn made by an airplane, or the slope of a curve on a highway or railroad. Increasingly **Web** and other **search** engines use automatic indexing to disambiguate or to create broad categories for browsing. 53

MAI can speed up the indexing and abstracting process needed to prepare databases. It particularly helps in handling such high volume tasks as assigning metadata terms to **Web** documents. 54

Automatic Summarization

Not too long ago, no one could find information. Now there is too much of it. Any tool that gets us quickly to the most important bits is valuable. Quick, automatically produced summaries have this potential. There are two kinds of automatic summarization. The first summarizes whole documents, either by extracting important sentences or by rephrasing and shortening the original text. Most summarization tools currently under development extract key passages or topic sentences, rather than rephrasing the document. Rephrasing is a much more difficult task. 55

The second process summarizes across multiple documents. Cross-document summarization is harder, but potentially more valuable. It will increase the value of alerting services by condensing retrieved information into smaller, more manageable reports. Cross-document summarization will allow us to deliver very brief overviews of new developments to busy clients. We can expect some tools to do this within the next 2-4 years. 56

Cross-Language Retrieval

Research communities now span the globe. Researchers need to know

what goes on in their fields no matter what the language of the source, e.g., companies going global in scope and interest. Two approaches are in development. The first translates text from one language to another. The second maps words in the same language to a single coded concept, just as concept mapping does. Even rough wording or poor translation is adequate for cross-language retrieval. We can also use it for retrieving foreign language documents, even if we can't translate the documents perfectly. The combination of concept mapping and automatic summarization can deliver a rough gloss or overview of an article so that a researcher can decide whether to read an entire document. 57

Entity Extraction

Entities are names of people, places or things. As we all know, entities are often difficult to locate within a collection of documents because many variant terms may refer to the same person. For instance, "AT&T" may also be found as "AT and T," or "AT&T." "Marcia Bates" may appear as "Bates, M" or "Bates, Marcia," but should not be confused with "Mary Ellen Bates." President Clinton was once Governor Clinton was once Governor Clinton and still is Bill Clinton and William Jefferson Clinton, not to mention "the President." 58

Newer information systems develop lists of name variants so that all the forms of a name map to the same concept and will retrieve all the records, no matter which term appears in a query. These systems may also contain built-in lexicons with specialized terms and geographic name expansions, e.g., to include France when the searcher asks for Europe. System administrators should have access to the lexicons to add internal thesauri and vocabulary. They should also add new names or terms as they occur in new materials. NetOwl is one example of a product that extracts entities. For decades, LEXIS-NEXIS has used name variants in order to improve retrieval, but automated extraction and storage give this policy far more power. 59

Relationship Extraction

With extracted entities in hand, one can perform some interesting analyses across documents. For instance, one could find out who has met with whom over the timer period of the collection. This kind of **data analysis** requires that the system extract relationships among entities. Some systems can extract more than 60 different types of relationships, including some that describe time or tense and numbers. Natural language researchers have developed categories to describe these relationships. For instance. 60

* The ISA relationship defines who or what the subject is: "Gil Shahan is a fine violinist."

* The AGENTOF relationship describes who or what caused an event to happen or had causal relationship: "Increased ozone in the Southern Hemisphere causes severe sunburns."

Tools like KNOW-IT, developed by Woojin Paik of Solutions United, extract entities and **store** their relationships to each other. This involves a larger chunk of information than single words or even phrases, consisting of the subject, the object, and the kind of relationship they have to each other. That way, we know who initiates what action and what its effect is on whom. These tools would **store** Jim owes Fred as a different unit than Fred owes Jim. The system can create webs of relationships that might help to direct which bacteria were becoming drug resistant as a result of which antibiotics or to detect which drug traffickers work together. 61

As we have seen, words by themselves often do not suffice to establish meaning. If one can **store** the context, the syntax, and the unambiguous meaning of each sentence as a unit, one can build a good question-answering system. Tools like this can answer questions such as, "Who fired the president of Consolidated Widget Company?" 62

Chronological and Numeric Extractions

If a system can determine when and what event Was happened, or howlarge something is compared to something else, then it can answer questions such as, "When was Netscape bought by AOL?" or, "Find all the Widget companies that produce -more than 5 million widgets a year." With this kind of information extracted from its contents, the system can also construct chronologies of events. This may not seem earth shaking, since one might find a biography of a person instead of constructing one, but imagine the-possibilities if the system could reconstruct the development 63

of a competitor and then use that model to monitor news for emerging competitors before you have identified them.

Text Mining

Text-mining technologies differ from searching because they find facts and patterns within a **database** . In other words, text mining, looks at the whole **database** , not just a single document, and then extracts information from all the pertinent documents in order to reveal patterns over time or within a subject. These technologies perform some analysis on text in a **database** to present patterns, chronologies, or relationships to the user.

Librarians do **data mining** almost implicitly -- to them, information falls into patterns, groups, clusters, and hierarchies. While it may seem second nature to us, in fact, it is a rare talent. How can software accomplish the same thing? Well, it can't with any intelligence. But remember that language is made up of patterns; this fact lets us generate new but still under-standable, sentences. If you identify the clues that tell you, for instance, that something is a prediction, then the software can follow those same rules to find predictions, e.g., using terms like "by next year," "in 2010." Good text mining depends on the quality of the knowledge base on which it operates. If relationships, concepts, chronological information, and entities have already been extracted, then, the text-mining process can take advantage of this information and seek patterns within it.

Question-Answering Systems

We often lose sight of the purpose of information retrieval, which is usually to answer questions, not just retrieve documents. Question-answering systems look within documents or knowledge bases to find answers. For example, if you ask a question-answering system, "When was the Wye River Accord signed?," you will get an answer of October 1998, rather than a list of documents about the Wye River Accord, which may or may not contain the answer. Question-answering systems find the best matching answers extracted from within matching documents. If users need more information, they can link to the source documents.

Filtering, Monitoring, or Alerting

The difference between filtering and ad-hoc searching is that in searching, the **search** may change, but the **database** remains the same, while in filtering, the, **search** stays the same, but the data against which the **search** matches changes. Filtering only looks for new documents of interest. To set up a filter, the user creates, a profile or "standing query," which runs against any new additions to the **database** . The art of designing a standing query lies in creating a broad enough query to prevent the omission of important developments, while making it narrow enough to prevent too much information from flooding the user.

Like any other **search** technology, filtering or alerting depends on the quality of the **search** engine used. A **search** engine that can provide well-focused retrieval, preferably using some sort of disambiguation and concept extraction, will most likely catch related topics.

One of the major problems with any kind of standing, continuing query, or monitoring service is that the terminology in any field changes over time. So do a user's interests Yet, most of today's alerting services are static. Those who rely on profiles must make sure to update them regularly. As an example, my own 3-year-old alert on "information retrieval" returns very little of interest these days. Instead, I need to add **search** engines, **data mining** , text mining, filtering and routing, natural language processing, knowledge management, and many other new terms. Newer systems that incorporate some kind of machine learning or intelligent agents are vital for good continuing monitoring of topics. Filtering tools that incorporate machine learning can detect new terms and offer to add them to a standing query. They can also, note changes in the user's interests and adapt the query to fit these new topics.

Change Monitoring

Change monitoring is a specialized type of filtering. It monitors established documents or **Web** sites and determines when changes have occurred within them. The technique has become a vital part of competitive intelligence or events monitoring. If a competitor' **Web** site remains unchanged, the system ignores it, but it raises a red flag if substantial changes and additions occur. Similarly, official agencies charged with collecting and archiving government documents need to know when a new

revision of a form or document or law appears.

One company that monitors **Web** pages for changes is Ingenius Technologies (http://www.ingetech.com). Their JavElink monitors a list of URLs supplied by the client and reports only the changes. The visual display makes it easy. to note what has changed at a glance (see Figure 4 below). Ingenius also uses this technology to create emailed alerts (NetBrief, http://www.netbrief.com) that contain only the changed text of a site; The Ingenius site displays several free alerts on popular topics as examples. 71

A new extension of NetBrief sends a daily e-mail containing URLs and brief excerpts matching client keywords. Each day, InGenius reviews 100 online daily newspapers, as well as dozens of **business** and technical publications. Clients may add new sites or **search** engines as they wish. They may also specifically include or exclude certain sources or topics. 72

Visualization

The human eye understands visual representations, much faster than it can read-text. As? the old proverb says, "One picture is worth a thousand words." Compare the simplicity and speed of recognizing a picture of people sitting under a tree at a picnic to reading a description of the same. scene. In order to help people interpret large sets of data or documents, many researchers are designing visual equivalents of the text, so users can digest the information at a glance. 73

Visualization helps handle information overload. Imagine being able to Hand a one-page visual overview of the week's developments to the CEO of a company instead of a five-page digest. Visual information systems are also vital to crisis management, air. traffic control, and other situations in which people must respond instantly to a great deal of information. 74

Effective Visual representations are confined by the limitations of-the computer screen. There is, only so much information that can be displayed effectively on the standard 14-or 15-inch monitor. For an example of a nice. kind of interface to have see a description of the interface to Phrasier (http: //www.cs.waikato.ac.nz/-stevej/Research/Phrasier), an innovative system for browsing. by phrases. The screen design for this product is too large to fit a standard screen, but it contains all the elements that a user would want to have in order to interact well with an information system. It displays documents, related concepts, and key Phrases, all in one place. Figure 5 below shows part of the screen. 75

Most of the visual presentations of information we see today are experimental. We really don't know how people will interact with them. Cognitive psychologists, online experts, and computer scientists need more than the anecdotal information we get front us-ability tests in order to establish guide-lines for, good design. We do know that people have many different cognitive styles and that to interact with computers efficiently they need tools and interfaces that fit how they think. The great challenge will be to discover how the mind works and then to design tools based on this knowledge. 76

Some concepts are fairly simple to visualize effectively. Bar charts or even differently sized squares can illustrate quickly comparative sizes, amounts, or numbers. Timelines can show time-dependent events. Proximity of objects can indicate close relationships. Pie charts show how the parts make up a whole. When we move from these common concepts to representing relationship among people and places over time, then we must invent new imaging. 77

A visualization sits on top of the information retrieved from a system. While the interface determines how the information displays, what it displays depends on the data extracted. Thus, relevance rankings easily display as bar charts. The amount of information available on a topic can show as a set of colored boxes of various sizes. 78

The vector space model that we discussed earlier lies under most visualizations of subject content. It can create star charts, showing clusters of documents, or the imaginary land-form maps from Cartia). Look at this visualization of a set of **search** results from Cartia in Figure 6 below. The highest peaks represent subjects having the most documents. The closeness of hills shows proximity. 79

The browser from the Human Computer Interaction Laboratory (HCIL, http://www.cs.umd.edu/hcil/ndl/ndldemo/draft11/daveloc4.html) at the University of Maryland gives an instant overview of the Library of Congress collections. As you pass your mouse over each timeline, it turns blue, and 80

so do the types of collections that contain information about that time period.

Query formulation is one of the weakest spots in the information process. Several companies and research groups have developed visual aids to query formulation, but I still like the text power **search** screen from DR-LINK, developed by Dr. Liddy at Syracuse University, that shows you how the computer has interpreted your **search** and gives you a chance to change it (see Figure 7 below).

Spotfire (see Figure 8 below) and Dotfire, its newest form, are dynamic query tools. These tools present a set of categories that help to narrow down a **search** . You can manipulate each category using a slider. HCIL at the University of Maryland developed both of them (http://www.cs.umd.edu/hcil). Dotfire (http://www.cs.umd.edu/hcil/west-legal/dotfire.gif) is the new West-law case law explorer. (For more information, read the technical paper by Ben Shneiderman, David Feldman, and Anne Rose, "Visualizing Digital

Library **Search** Results with Categorical and Hierarchical Axes," CS-TR-3992, UMIACS-TR-99-12, February 1999, ftp://ftp.cs.umd.edu/pub/hcil/03html/99-03.html.)

Figure 9 above shows the Hyperbolic browser from Xerox PARC, developed to help people explore the contents of a **database** visually You can find it at the InXight **database** . (http//www.inxight.com).

Gary Marchionini and his students at the at the Interaction Design Lab at the University of North Carolina study the effectiveness of interface designs for various kinds of resource formats, such as statistics or video files. The interactive statistical relation browser is a pro-totype developed for the Bureau of Labor Statistics. It displays, in one screen, subjects covered by the **database** , the number and format types for reports, as well as regions and dates covered. Related **Web** -sites also display. It is simple, effective. (See http://ils.unc.edu/idl/ for other research by this group.)

The Perspecta interface shows the user, in one screen, which parameters they can **search** . This screen shot also shows the results of a **search** done on their travel information **database** Each box shows the user, at a glance, the number of tours that exist in each of the categories requested during the time period indicated For instance, 87 canoeing tours are offered during a specific time By grouping results into logical bundles this software enables the user to understand the results of a **search** before be the to plow through actual hits:

Having tools that can give you several views of the same data helps you discover patterns.

Northern Lights custom folders give you a quick visual overview of **search** results The careful categorization of contents makes searching Northern Light both broad and well focused. Northern Light also searches Yahoo! directory pages. Yahoo! has some excellent resources, but I prefer to **search** rather than to start with a browse. Northern Light gives me the best of both approaches.

I like the simple display from TASC, (www. tir.tasc.com/Visualization/). TextOre shows the extent of the information about a subject by the size of the colored squares. If you click on a square, you will see the documents that it represents, or, for large document sets, further charts. This is visual **data mining** .

Tools to Analyze and Interact with Data

Finding and using information should be an active process. We need to read what we find, but we also need to merge sources, pull them apart, separate the data, into categories, sort the data, seek patterns, and send the information to colleagues and clients.

Puffin **Search** (http://www.puffin-ware.com) invites this kind of interaction. It searches across up to eight **Web** **search** engines at a time and brings the results back to your desktop. It saves the **search** results, creating a list of all the terms that appear in two or more citations. Then you can sort, cluster, and resort the results using any cell in the table as a basis for comparison. Choose a title and it will re-rank all other hits by their similarity to that title. Or, choose several of the keywords and rank all 1,200 hits by the terms you have chosen. You can sort by **search** engine or by URL. Puffin automatically forms clusters based on the similarity of a group of documents, using a similar technique to the vector space model. You when you use it as a

filtering tool.

Netbook, developed by the Human Computer Interaction Group at Cornell University (http://www.hci.cornell.edu),is part of a multimedia tools suite that foreshadows what the digital library will look like in the future. These are the tools that users will demand as we move to, dynamic use of information (http://www.hci.cornell.edu/projects/projs/multimedia.htm): 92

Netbook allows users to capture images from digital collections, **store** , and view them on a user's personal netbook page. Thumbnails of images can be organized, allowing users to build a manageable collection. 93

Annotator allows users to annotate or read annotations to images that have been taken from an online collection or a Netbook. A class of students for example can view the annotations an instructor has made to an image or create their own annotations visible to their classmates. 94

Authoring helps users manipulate, organize, and display research and data with hierarchical and hypermedia links. Students can view the work of others and organize andmake their own links. 95

Artview transforms online image collections from museums and other sites into collaborative learning spaces. At the same time but from different locations, users can view an image and communicate with each other using a shared text window. 96

Searching Multiple Sources Simultaneously

Searching across different kinds of information collections poses one of the biggest challenges facing digital library and intranet builders. Collections may encompass text or images or statistics. Text files may contain bibliographic records, abstracts, or full text. Image collections may only offer **search** engines the text appearing as captions. Once we move outside of controlled, integrated collections of the same kind of Materials, we encounter several obstacles. These include vocabulary differences, differences in type of materials, and differences in relevance-ranking algorithms. 97

Differences in vocabulary are a familiar problem to .any experienced searcher; Each collection or source may use different terms to express the same idea. We professional searchers traditionally handle this problem by using every synonym we can think of. Thus, we might choose both pumps and impellers, or theater and theatre to round out a good query. In NLP systems, concept matching may perform some of this work for us. However, customized intranets may want to develop internal lexicons that would map pumps and impellers to the same concept automatically This is a good application for concept monitoring and indexing. 98

Searching across heterogeneous materials presents a Knottier problem, as searchers working with Dialog OneSearches can tell you. For instance, the weight of each word in a bibliographic record is probably enormously high compared to the same term appearing in a full-text, 10-page 'document. One could imagine trying to tweak a **search** system each time it adds a new kind of collection. 99

Searching across several systems complicates matters still further. Most **search** engines calculate the relevancy of a document by counting the number of occurrences of each query term in each document. The more occurrences, the more relevant the document. This works fine when the documents are approximately equivalent in length and of the same type. When we combine these materials in a single **search** , the results will skew by length of text. 100

If we try to **search** across **search** systems, as **Web** metasearch engines do, we find that each one measures relevance differently. In addition, since each system computes the relevance of a document to a query in part by finding out how rarely that term occurs in the **database** as a whole, and collection contains different materials, it is unlikely that what is highly relevant in one collection will rank the same way in another. Data fusion is a set of techniques for establishing a common ground to measure relevance. The lack of data fusion treatment explains why Searching across files in Dialog or metasearching on the **Web** doesn't work well within relevance-ranking systems. 101

Here's an example. Suppose that: we decide to **search** for a few good articles on the causes of high blood pressure. We pick two **Web** **search** engines. But, we don't know that **Search** Engine 1 covers all the major medical information sites, while **Search** Engine 2 concentrates on sports. **Search** Engine 1 finds 250,000 articles about high blood pressure. It ranks them. **Search** Engine 2 finds 10 articles, and they have only minimal 102

information on the subject. Think back to our weighting algorithm. If high blood pressure appears rarely in a **database** , it gets a high weight. SO, **Search** Engine 2 gives all of these documents a 98 percent ranking. Since high blood pressure constitutes a common term in **Search** Engine 1, it gets a lower weight. If our metasearch engine takes the top 10 from each, we will see all 10 of the **Search** Engine 2 documents before we ever get to those from **Search** Engine 1. Yet, the results from **Search** Engine 1, coming from medical sources, may be vastly superior.

Data fusion tries to merge results from several **search** systems. One technique takes one document from each in a round-robin approach. Another creates a virtual collection that merges all the documents found in all the databases. Then weights are reassigned based on this common collection. The second technique gives better results, but is computationally more costly. 103

Evidence Combination

Evidence combination improves retrieval from the same collection by using different retrieval techniques. It will be a hot topic in the next few years, as computing power increases still further. Any retrieval technique is faulty and will omit some relevant documents, perhaps due to a poor query, to differences in terminology, or even to errors in spelling introduced by optical character recognition programs. Searchers may also miss important documents if the documents do not appear in the top 30 or 50 examined. Certain ranking algorithms clearly do a better job on one type of document or another. Some may adjust for word position or proximity of query terms. Others are partial to long or short documents or tend to give priority to term frequency instead of to term rarity in the **database** . Some may emphasize metadata; others ignore controlled vocabulary terms entirely. These are all reasonable design choices that may conform to a particular type of collection. While searchers cannot always understand why one **search** engine misses certain documents that another retrieves, we know that this happens. The differences in **search** algorithms may offer one explanation. 104

Evidence combination can refer to searching the same collection with different **search** engines and combining the results, or it can refer to using different sources to gather information about documents. For example, a collection of newscasts might be searched from speech text created by speech-recognition software. Closed-caption broadcasts would supply another source, and so would the video images themselves using image-recognition software. Each one of these sources is not a reliable source by itself - none of them contains enough accurate information on the subject of the document - but combined, the strengths of one make up for the weaknesses of another. Informedia (http:// http://www.informedia.cs.cmu.edu/), one of the first National Digital Library Projects, offers a good example of this technique. 105

Speech Recognition for Spoken Interfaces

Although we have become reasonably comfortable interacting with the computer by keyboard and mouse, it is not natural. Our interactions show it. Who would ask a spoken question with a single word? Yet, the vast majority of queries on **Web search** engines are single words. And would we really choose to input a query with parentheses and truncation symbols, given a simpler alternative? Spoken interactions are a more normal mode and a voice interface, or VUI (voice user interface), may solve some of the input problems that designers face with written or graphic interfaces. 106

There are two distinct sides to voice recognition: input and output. Speech recognition can go from text to speech or from speech to text (speech synthesis). Both speech recognition and speech generation software must be developed in order to create good VUIs. The easier of these is speech generation. People already can understand computer-generated speech because they already know how to adjust to slight variations in pronunciation or intonation, if only from listening to real people speak. Companies like Cogentex have already created technologies that generate speech from data plus a template. The Montreal weather report uses this product.

Voice recognition is a more difficult proposition. Natural-language processing gets us part of the way to voice-recognition systems, but a few levels of language important in speech make problems in written language. The way we pronounce words has many more variations than we realize. For instance, the "c" in "cat" differs from the "c" in "core." Intonation patterns convey meaning by the song that is sung. A declarative sentence 107

versus a question, for instance, is solely distinguished by the notes that the voice uses - a falling instead of a rising inflection. Voice recognition also stumbles on regional pronunciation differences, as well as on finding the boundaries between words. We run one word into another and expect our listeners to make the cut between each one. Computers can't manage this as easily. Try saying, "What's to stop me?" in a normal tone to see what I mean. Gotcha!

Nevertheless, voice interfaces have begun to appear. MyTalk (http://www.mytalk.com) from General Magic will fetch your e-mail and read it to you on the phone. It uses speech generation software and intelligent agents to read only what you want. You can interact using several hundred commands and, if you forget what to ask, it will give you choices. 108

Microsoft, with its SAPI standard (Speech Application Programming Interface) Persona Project, and associated Speech Recognition research groups, seems to be creating a successor to Microsoft Bob, which can interpret continuous speech and then generate an answer. Microsoft uses NLP and could apply this software to information retrieval as well. Other major players are Lernout and Hauspie, Dragon Software, IBM, Nuance, Motorola, Unisys, Dialogic, and AT&T. 109

Most of the research with NLP and speech recognition concentrates on understanding word boundaries and correctly identifying phonemes across diverse speakers and accents. This is a non-trivial task. One solution is to train a system within a small domain, such as answering customer-service questions for one particular company. Another is to train an application to recognize only one user's voice. This latter application is in demand for those who can't read a screen or type. In fact, reporters with carpal tunnel syndrome form a growing group of VUI users. 110

Once we solve the problem of establishing normal speech interaction as a computer interface, our whole mode of operation with computers will change. We will ask our car for directions and have it tell us where to turn next, after it has mapped out our route. In fact, that is available now. We will tell our agent to read us any news on **Internet** -related subjects while we make the coffee. It will ask us if we want to hear the urgent message from our boss first. And, we will ask for the monthly report to be generated from statistics and then presented as a PowerPoint presentation, complete with pie charts, without having to remember how to import a chart and resize it. 111

It's nearly 2001. Can HAL be far away?

Designing the Answer Machine

* Study users and their needs.
* Design the output and then the access system.
* Design the input to create the output you need. 112

Know your users and their work situation. Are they a captive audience? Can you offer training and will they take it? What kinds of information do they need and in what formats? Do they need in depth analyses, research reports, top-level summaries, weekly briefings?

Researchers look for information differently from marketing people or executives. This isn't surprising, since they all have different kinds of information needs. Researchers want in-depth information. Marketing people may want facts, statistics, or to keep up with the competition. Executives may want quick overviews and summaries that give them a lot of information at a high level in a capsulized form. Do your users want everything on a topic (high recall) or just the few best nuggets high precision)? How will they use the information? Do they need 24 hour, 7-day-a-week access from remote locations? What should the system output look like? 113

First, find out what your users need, want and how they will use the information. Then, design an access system that fits how they think and work. For instance, we have never found a user population that can distinguish between "subject headings" and "keywords." Don't expect that they will learn. Just create a system that doesn't require too much knowledge unrelated to their day jobs. 114

Create access models--subject, author, fields -- that make sense to your organization, even if this goes against library orthodoxy. If you only have computer science materials, don t expect that the Library of Congress Classification will be useful. Think about why classification schemes were invented and then use something that can help distinguish among the materials. 115

Last, design the system so that it will give you what you have 116

already specified. Don't get talked out of important features. And don't let fancy bells and whistles that will confuse the users to creep in. Keep it simple. Make sure that it is easy to navigate. Test it and retest it.

I am not necessarily a fan of all things automatic. The best systems give users an opportunity to interfere, add information, alter directions, and make corrections. These systems form a partnership with the user. When designing an information system, include the user in the design In the best of all possible worlds, system designers would observe how people use information within the work place and then design a system that fits into the normal work flow. 117

Conclusion

All these technologies add up to a seamless suite of information tools that will find information, organize it, keep it up to data, forage for patterns, and present understanding. In other words, an Answer Machine. The tools I have just described will enable us to understand large and complex sets of information more easily. These tools enable quick understanding by adding a new dimension of analysis and even fun to working with information. They will give knowledge workers the ability to examine, manipulate, and understand the information we retrieve for them. Using these tools, we can move up a level of abstraction to analyzing evaluating and planning. This will offer our profession an exciting, challenging role bright with promise. 118

To be involved in the development of the next generation of information system, we must be willing to think big, stepping back occasionally from deadlines and from gathering isolated facts and statistics. We must comprehend and clarify the place of information in the organization. This is a role for practical visionaries. 119

Fortunately for us, that's exactly who we are. 120

Susan Feldman is president of Datasearch, and of Datasearch Labs, a usability testing company for information products. She writes frequently on new information technologies, and tests, evaluates and recommends products for clients. 121

 122

Here are some samples of questions difficult to answer in traditional information retrieval systems:

Identify bacteria in the process of becoming drug resistant.

Identify Bermuda advertising campaigns that promote the island as a tourist attraction.

Provide articles and case studies on attitudes of companies towards media relations, including best practices for approaching the media and trends in media relations.

Provide information on "issues preparedness" (i.e., rationales for why companies should be prepared to manage a crisis or issue in advance and how companies can effectively manage a crisis or issue).

Provide information on "thought" retreats/seminars/executive meetings, CEO retreats; and customer entertainment/ appreciation events.

Identify books or articles that discuss how artworks through the ages have represented oral hygiene and dentistry (for example, is there a reason why the Mona Lisa doesn't smile?!).

Identify emerging competitors in X industry.

Where should I go for my vacation in January if I don't want to spend more than $600 per person and I don't like crowds? I'd like to go some place warm with nice scenery, somewhere near an ocean.

How many widgets will Zambia manufacture in the next 5. years? I just want a number for each year, not a pile of documents. I need this in a half-hour, by the way.

I need to keep upon new information technologies as they appear. (This means that I need to identify new terms and also to drop those that have become outdated.)

Tell me when my competitors have come out with a new product. I don't want any other press, releases.

WHAT EXACTLY DO MEAN BY MEANING?

People extract meaning from text on many levels:

* Phonetic is the actual sounds made when we pronounce words. This isn't pertinent to written text, but it does convey extremely important shades of meaning in speech.

* Morphological is the smallest unit of language which conveys meaning. This includes plural versus singular forms, as well as other prefixes and suffixes, like pre- or -ization.

* Syntactic is the role each word plays in a sentence. Many of today's **search** engines can parse a sentence, as we learned to do in elementary school, in order to pick out the subjects, verbs, objects, and phrases. This enables the engines to distinguish between Bill picked Al and Al picked Bill.

* Semantic is the dictionary meaning of a word, as well as the meaning of a word supplied by its context in the text. This level helps us to distinguish the difference in meanings of "pool" in "Let's play pool" and "Let's swim in the pool." This ability to distinguish among the many senses of the same word is called disambiguation. It enables a system to eliminate false drops. An NLP system should never give you financial institutions if you ask for erosion of river banks, not even for the Consolidated Bank of Moose River.

* Discourse is the structure of a whole document. Many documents have a predictable structure, such conclusion. Where a sentence is placed in this structure influences its meaning and its importance.

* Pragmatic is knowledge of the real word. For instance, when we say Europe, we know this geographic region includes France, even if the document never explicitly states this fact. This kind of knowledge can be added to new information systems so that the systems understand that Congressman Schumer and Senator Schumer are the same person.

(For a more extensive discussion of NLP, see Sue Feldman's article, "NLP Meets the Jabberwocky,"
http://www.online.com/onlinemag/OL1999/feldman5.htm, Online, May 1999.)

BEFORE AND AFTER

Smith Widget, Inc., October 5, 1999. 10:00 AM

Boss: Good Morning, Dennis. We need to update our competitive intelligence report today I'd like to know all the new products our competitors have come out with in the last 6 months, as well as any plans they have for new products.

Dennis: Okay. When do you need it? The president just called me and wants figures for the board meeting by noon today.

Boss: Well, I really did want it by noon too. It's also for tomorrow's board meeting. See what you can do.

Dennis: I'll do my best. What information do you need the most? I'll work on that first.

Boss: Well, I really need a list of new products and their sales figures listed by company, and then I want a summary of trends and predictions for the industry, just bullet points.

Dennis: I think I can get the list of products for you, since we already know. the names of the companies, and I've been keeping a file of the changes to their **Web** sites. At least we have the new product announcements. It'll take a while to wade through the documents I get from an online **search** though, so I'm not sure I can get you the other information right away. I'll do my best.

Boss: I really need them in a hurry so we can get the graphics people to turn them into a slide briefing.

Dennis: When is the board meeting?

Boss: Tomorrow at 1 PM.

Dennis: I can probably get you the bullet points tonight and give them to graphics for tomorrow morning.

Boss: Well, if that's the best you can do, I guess we'll just have to settle for it, but I did want to review the notes tonight.

Dennis: I'll see if I can give you some preliminary results by 5 today and then work on a summary and bullet points. We can give the sales figures to graphics as soon as I get them. They're the easier part.

Boss: Okay. Just let me know as soon as you have something.

Dennis: Okay. (Boss leaves, Dennis dials wife's office). Hi. Guess what? It's quarterly panic time again. He wants a report by tonight. Can you call the Groves and ask if we can reschedule that dinner? No, I don't know the number of a good divorce lawyer, and I'm going to have a long enough day without any sarcasm. You know I love you. Sure, honey, see you

when I see you.

Dennis (musing): Now where did I **store** that CI **search** strategy? Okay here's Dialog, here's NEXIS, here's Dow Jones. I'd better update the **Web** filter, too, and look at those documents in my widgets CI mailbox. Here are the strategies. Dialog, file 16:ss (Jones or Franklin or Thomas or Automated) (w) Widget? and (ec=65? or ec=33?)

Search 2:ss pc= and (ec=1? or ec=6?) and (predict? or projecting or projected or future or forecast? or trend or outlook or year0 (200? or 201?)

(2:30 that day)

Dennis (calling boss): Hi, I have the product info and sales figures for your three competitors. Shall I send them to you electronically? There were 467 documents from the online **search** , and I'll scan them as fast as I can to get you the info you need. I'm using Puffin **Search** to merge and relevance rank the searches I did in Dialog, NEXIS, and Dow Jones. Is it okay with you if I just start with the top 150?

Boss: Yes, but please try to scan the rest too. We really got into trouble when we missed that new company, Automated Widgets, last time. I think they are marginal, but it doesn't hurt to see what they're up to.

Dennis: I'll do my best, but the last train leaves at 9:30, and I have to catch it.

Boss: Well, give me what you have by 9:00.

OUTCOME: Dennis had to quit, having missed both lunch and dinner, at document 322, in order to have time to write the summaries and bullet points in time. Document 463 showed that Automated Widgets had hired an expert in networked appliances from Sun Microsystems. Smith Widgets was bought out by Automated Widgets in 2003. Boss took early retirement. Dennis went on to help create a company-wide information system, designing templates for interaction and categories for automatic indexing.

Automated Widget Company, October 5, 2009. 10:00 AM

Boss: Good Morning, Alvin. We need to update our competitive intelligence report today I'd like to know all the new products our competitors have come out with in the last 6 months, and any plans they have for new products.

Alvin, the Computer: Okay, boss. Do you want products from your competitors if they are in a different product category from Automated Widgets?

Boss: Yes.

Alvin: When do you want this? What format?

Boss: I need it by noon today. Give me lists of products, organized by name of company. Then I'd like of summary of trends in the industry. Just summarize and make some bullet points, but keep the information. I may want more details on some of the major points in the summary. We're really worried about MS Widgets, so give me everything you can find on them. I want recent hires and firings, and any industry analyst reports.

Alvin: Do you want sales figures like the last report?

Boss: Oh, yes. I want sales figures for each. Compare them to the figures we have for that company 6 months ago. Just pull the old chart out of the last report and add a column for the new products and another for the sales. Also, give me any growth or decline in overall sales for each company. Don't forget their previous products.

Alvin: Anything else?

Boss: Yes, After you get me the lists, and the bullet points, update that competitive intelligence report we did 6 months ago.

Alvin: Same format?

Boss: Yes, but make the charts a larger font size. Also, extract the major points and put them and the charts in a slide presentation. Give me a separate slide on MS Widgets. I want that one by 1 PM.

Alvin: Anything else?

Boss: No, that's it.

Alvin: Okay. I will find lists of Automated Widgets competitors and their new products with sales figures and produce a list for each company. Then I will find trends and predictions. I will extract major points that appear in two or more articles or are mentioned several times in one article. I will deliver these lists and bullet points by noon to your inbox.

I will update the CI report from March 31, 2009, and use the new major points and charts for a slide presentation. This can be ready by 1

PM, but cannot be printed by then. The marketing department has the color printer reserved all afternoon. Can you review the slides online, or should I notify the printer that your work takes priority? We can print after 4 PM.

Boss: I will review online. Make the print font big enough to read.

Alvin: 14 point type font?

Boss: Okay.

(11:30. Boss walks into room)

Boss: Alvin, is the report ready?

Alvin: It is ready, boss. Printed copy is in your inbox. Online copy is in the high-priority info box labeled competitive intelligence. Do you want me to read it to you or do you prefer to view it?

Boss: Read me the new products and major bullet points. Also anything you found that doesn't fit a category

Alvin: In the new product category;

Franklin widgets Programmablerefrigerator/stove module

MS Widgets Programmable bathtub module

Widgetech Programmable gas grill

Programmable clutter hider

In the people category, Andrew Wyatt gave a talk in September at the Futuretech conference. I summarized it for you. You met him at the WIA conference last spring, and I have a note to tell you to contact him in October. His phone number is 577-304-8976. His e-mail is aww@futuristics.com. I have his street address, too.

In the unpleasant surprises category, you didn't ask to monitorSolutions.com. It is a new company that matches your competitive profile. They have developed a "company's coming" remote control module that hides clutter, inventories the refrigerator, orders groceries, cleans the house, turns on the oven, and changes the sheets.

In the MS Widgets report, their earnings have gone up 23 percent. They have just acquired a widget integrating company.

Is there anything else you want?

Boss: Yes! Get me everything you can on widget integration companies. I want a list of those with actual products, and what those products are. Also, sales and predictions for each of them.

Add Solutions.com to our monitoring list.

Alvin: Okay boss.

OUTCOME: Automated Widgets is slugging it out with MS Widgets at the moment. Will either of them notice Solutions.com sneaking up on them? This is a case of dueling information systems. Winner take all. Which one has the better technology for raising red flags? Which do you think?

IMPLICATIONS FOR INFORMATION PROFESSIONALS

The dawn of a new era can be exciting or unsettling. Right now, there are so many fingers in what used to be our information pie that we may reel crowded and, perhaps, threatened. Computer scientists, psychologists, graphic designers, linguists, and **Internet** businesses are all carving out pieces for themselves.

What do we information professionals have to offer of value? First, we have a unique perspective about information itself. We understand how to ask the right questions in order to kind what we need. We understand balance in collections, good sources, and how to categorize materials so people can find them. This is invaluable. We also have something the others may lack -- we use information systems. We have searched for information for decades. We have practical experience. If we can temper the experience with the flexibility to try something new, we can become the part of the development team most firmly anchored in reality.

Things brings me to some tentative ideas on what to look for as you go about putting together an intranet or information system for an organization. These thoughts are tentative because they haven't been tested, and may theories are just as suspect as anyone else's. I can only rely on my own experience and tests of technology. Based on my comparisons of NLP systems with other systems, I know that NLP systems work and work well. Similarly, I have been extremely pleased with the agent systems and automatic indexing systems with which I have experimented. So, I know that the foundation technologies work and much better than anything else I've tried. I think that if I we were putting together a system for tomorrow, though that I would look for products with these technologies as my base.

PUBLISHER NAME: Information Today, Inc.
EVENT NAMES: *360 (Services information)
GEOGRAPHIC NAMES: *1USA (United States)
PRODUCT NAMES: 7375000 (**Database** Providers); 4811520 (Online Services); 7372421 (DBMS); 3573025 (Document Processing Computer Systems)
INDUSTRY NAMES: LIB (Library and Information Science)
NAICS CODES: 514191 (On-Line Information Services); 51121 (Software Publishers); 334111 (Electronic Computer Manufacturing)
SPECIAL FEATURES: LOB
?

Save Link Saves this search as a Durable Link under "Results-Marked List"

At least 11 articles matched your search.

1. State of the Industry: Measurement Finally Coming of Age in 2003; **PR News**, Potomac; Feb 24, 2003; pg. 1
2. Trends & Tactics ...; **PR News**, Potomac; Feb 24, 2003; pg. 1
3. Biz360 and CARMA Form Strategic Alliance to Offer Comprehensive Solution for Corporate Executives to Understand Value of Communications Initiatives; **PR Newswire**, New York; Jan 28, 2003; pg. 1
4. Biz360 and Dialog Form Alliance to Help Corporate Executives Manage Their Company's Perception in the Media; **PR Newswire**, New York; Jan 28, 2003; pg. 1
5. PR Newswire Northern California Summary, Tuesday January 28, 2003 Up to 2:00 p.m. PT; **PR Newswire**, New York; Jan 28, 2003; pg. 1
6. Trendwatch; *Alison Stateman*; **Public Relations Tactics**, New York; Dec 2002; Vol. 9, Iss. 12; pg. 3, 1 pgs
7. Biz360 Introduces Objective Media Measurement for Public Relations in the Pharmaceutical Industry; **PR Newswire**, New York; Nov 20, 2002; pg. 1
8. Biz360 CEO to Speak on Total Reputation Management at the PRSA International Conference in San Francisco; **PR Newswire**, New York; Nov 13, 2002; pg. 1
9. PR Newswire Northern California Summary, Tuesday November 13, 2002 Up to 2:00 p.m. PT; **PR Newswire**, New York; Nov 13, 2002; pg. 1
10. Perceived reality; *Erin Strout*; **Sales and Marketing Management**, New York; Nov 2002; Vol. 154, Iss. 11; pg. 26, 1 pgs

Next ▶
11-11

Refine your search. Enter a word, words or specific phrase.

biz360 Search

Date range: Current (1999 - Present)

Publication type: All

Search in: Article text

- [x] Show results with full text availability only
- [ ] Show articles from peer reviewed publications only
- [ ] Show total number of articles

Save Link Saves this search as a Durable Link under "Results-Marked List"

At least 11 articles matched your search.

1. Metro Atlanta Chamber Selects Ketchum/Crescent As Communications Agency for 'Industries Of The Mind' Initiative; *Business Editors*; **Business Wire**, New York; Dec 31, 1998; pg. 1
2. Becoming Market 'Priya'; **Dataquest**, New Delhi; Dec 31, 1998; pg. 1
3. Mid-year Review; **Dataquest**, New Delhi; Dec 31, 1998; pg. 1
4. WIRED INDEX OF `NEW BLUE CHIPS' REFLECTS THE INFORMATION AGE; [Broward Metro Edition]; *HUMBERTO CRUZ Columnist*; **Sun Sentinel**, Fort Lauderdale, Fla.; Dec 27, 1998; pg. 7.F
5. After a rosy 1998, retailing looks to a weaker new year; *Susan Chandler*; **The Charleston Gazette**, Charleston, W.V.; Dec 25, 1998; pg. 5.D
6. BYTES; [FINAL Edition]; **The Washington Post**, Washington, D.C.; Dec 21, 1998; pg. F.05
7. RETAILERS CELEBRATE WHILE GOOD TIMES LAST; [CHICAGOLAND FINAL Edition]; *Susan Chandler, Tribune Staff Writer.*; **Chicago Tribune**, Chicago, Ill.; Dec 20, 1998; pg. 1
8. Don't Bother Making a Federal Case Out of Break Time; [FINAL Edition]; *Kirstin Downey Grimsley*; **The Washington Post**, Washington, D.C.; Dec 20, 1998; pg. H.04
9. Capitalizing on the Internet; *Anonymous*; **National Journal**, Washington; Dec 19/Dec 26, 1998; Vol. 30, Iss. 51/52; pg. 3014, 4 pgs
10. MALL VS. MOUSE;; [Broward Metro Edition]; -- *VICKI McCASH Assistant Business Editor*; **Sun Sentinel**, Fort Lauderdale, Fla.; Dec 19, 1998; pg. 1.D

Next ▶
11-20

Refine your search. Enter a word, words or specific phrase.

mindshare Search

Date range: Backfile (1986 - 1998)

Publication type: All

Search in: Article text

- [x] Show results with full text availability only
- [ ] Show articles from peer reviewed publications only
- [ ] Show total number of articles

ents

San Jose Mercury News: Egypt Building Monument To Tech

Christian Science Monitor: Ancient Egyptian library reborn in modern form

New Scientist: Way Back When

NPR: Library for Kids Goes Online

NEW! Compare Archive Pages with DocuComp®

O'Reilly/Koman on the Bookmobile and the Public Domain

Slashdot: Public-Domain Bookmobile Hits the Road

Internet Archive Bookmobile Launch Party

Library of Congress Acquires Prelinger Collection

Donation to the new Library of Alexandria in Egypt

ews

Marc Broussard: 2002-12-31 Great show!

Sound Tribe Sector 9: 2003-02-08 A Hot Time in the Olde Towne Tonight

Archive Collections

The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.

INTERNET ARCHIVE
WayBackMachine

http://www.biz360.com

Take Me Back

Advanced Search

The Internet Archive, working with Alexa Internet, has created the Wayback Machine. The Wayback Machine makes it possible to surf more than 10 billion pages stored in the Internet Archive's web archive. The Wayback Machine was unveiled on October 24th, 2001 at U.C. Berkeley's Bancroft Library. Visit the Wayback Machine by entering an URL above or clicking on specific collections below.

Browse the Internet Archive

ICDL

The International Children's Digital Library where kids all over the world can find lots of books from many different countries.

The Internet Bookmobile has gone from SF --> DC to celebrate the public domain! Check in and see the voyage and meetings with students, other bookmobiles, and librarians. More ...

Moving Images

The Internet Archive is collaborating with various collectors, community members, and film-makers to provide easy access to a rich and fascinating core collection of archival films.

- **Prelinger Archives**
- **Computer Chronicles**
- **Net Café**
- **World at War**

Texts

The Internet Archive is collaborating with numerous libraries to digitize as many texts and books as possible.

- **Project Gutenberg**
- **Million Book Project**
- **UVA**
- **Liber Liber**
- **Arpanet**
- **Open Source Books**
- **Internet Bookmobil**
- **Internet Children's Digital Library**

Audio

etree.org

The Internet Archive is collaborating with etree to provide the highest quality live concerts in a lossless, downloadable format.

- **tr e Audio Archives**

Software

Macromedia and the

Entrepreneur Profile: YOU MO[redacted]TSANG

San Francisco Business Times; San Francisco; May 31, 2002; Petra Pasternak;

*NAICS:***511210**
Volume: 16
Issue: 43
Start Page: 38
ISSN: 08900337
Subject Terms: Personal profiles
Chief executive officers
Software industry
Classification Codes: **9190**: *United States*
8302: *Software & computer services industry*
2120: *Chief executive officers*
9160: *Biographical*
Geographic Names: San Mateo California
Personal Names: **Tsang, You Mon**
Companies: **Biz360** *NAICS:***511210**
Full Text:

RESUME

Name: You Mon Tsang.

Title: CEO and founder.

Company: Biz360 Inc., in San Mateo; develops software that helps companies measure the impact of their PR and marketing departments.

2001 revenue: $1 million.

Number of employees: 35.

Year founded: 2000.

Source of startup capital: Seed financing from founders, later investments from Granite, Adobe Ventures, and Foundation Capital.

Background: B.A. in urban studies from Yale University, MBA from Haas School of Business. After working as an urban planner, became director of product management at Brio Technology and was founder and CEO of Milktruck.

Age: 36.

Residence: Burlingame.

Web site: www.biz360.com.

BIG PICTURE

Reason for starting business: It's what I do. If it wasn't Biz360, I'd open a restaurant or a design studio. A better question to ask me is, "Why not start a company?" It's natural for me. I love the tasks of growing a business. I love coming up with ideas and actually putting together great people to solve problems or provide a service.

Hardest part of the decision: How to most gracefully leave my previous job at Brio Technology. I just really enjoyed the

company and the challenges I had there

Biggest plus of ownership: There's often this misconception that you have a ton of control as an owner - like the driver of a speedboat it turns left or right the second you make the decision. I'm (more like a) tugboat captain that nudges a huge ship to an ultimate goal - and thats what it's like to run a growing business. I like setting the direction.

Biggest drawback of ownership: I love what I do, what we built as a company, and I expect that from everyone who works for me, which is, frankly, unrealistic. When I don't see it, it's hugely frustrating.

Biggest misconception about ownership: A lot of it is not under your control. You're dealing with the economy and other external factors.

Biggest business strength: That I have a good sense of what the next 18 months looks like.

Biggest business weakness: I need to step back and know that my employees don't need my help all the time.

Biggest risk: To identify a new way of doing business - a new product and service and creating a new marketplace without really knowing beforehand whether that market existed. It's going great We have 40 customers today (including) HarleyDavidson, VeriSign, Xom and Cap Gemini.

Biggest worry: These days - trying to manage growth at the same time that I'm watching the bottom fine.

Top source of inspiration: I have my best ideas when I'm at a concert listening to five music. Now I always have a pencil ready, writing on a cocktail napkin. And, my wife and kid. They recharge me. I am lucky enough to love to come to work in the morning and to love coming home at the end of the day.

DAILY ROUTINE

Most challenging task: Managing the growth.

Favorite task: Selling the company's vision and service.

Least favorite task: I hate getting on business time. I'm in at 8 a.m. and out at 7 p.m.

Greatest frustration: When I don't see people taking ownership of their tasks, their jobs or their customers.

Sources of support in business crisis: My board, the senior team, entrepreneurial peers, my wife and friends in other industries for a different perspective.

DREAMS

Goal yet to be achieved: An IPO.

First move with capital windfall: Today, I'd put it in the bank.

Five-year vision: As a very large company - anywhere from 500 to 750 people; a major corporation, public, and a leader in high tech.

Inducement to sell: A whole lot. I don't think anyone can match my expectations or optimism.

First choice for new career or venture: I probably have one more large company in me, and then I'd like to start a small no more than 25 people - graphic design company in San Luis Obispo.

PERSONALS

Most-admired entrepreneurs: Richard B[redacted]son of Virgin, and Herb Kelleher of S[redacted]hwest Airlines.

Most interested in meeting: Jimmy Carter.

Stress reducer: I chase my 18-month-old son around the table.

Favorite pastimes: Hiking, outdoor sports. I'm on a softball team.

Favorite book: "Common Ground: A Turbulent Decade in the Lives of Three American Families," by J. Anthony Lukas.

Favorite film: "Brazil."

Favorite destination: Pt. Reyes National Seashore.

Automobile: White Subaru Forester - it's my family car - and a black Mercedes 560 SL, my fun car.

Biz360's First Solution Measures Company Buzz

PR Newswire; New York; Feb 12, 2001;

*NAICS:*334119 *NAICS:*334419 *NAICS:*7373 *Duns:*09-995-6906
Start Page: 1
Dateline: Arizona, California
Companies: 3Com Corp *Ticker:*COMS *Duns:*09-995-6906 *NAICS:*334119 *NAICS:*334419 *NAICS:*7373

Abstract:
Biz360 is a provider of automated analytic solutions for external company information. Biz360 solutions measure information previously thought unmeasurable like industry buzz, company mindshare and technology trends. Biz360 solutions combine the power of business analytics with the breadth of Internet content to deliver completely new, real-time market information and analysis services. Biz360's first solution, Market360, is directed to Marketing executives, Public Relations departments and PR agencies. Biz360 is a privately held company located in San Mateo, CA and is funded by Granite Ventures (formerly Hambrecht & Quist Venture Associates) and Adobe Ventures. SOURCE Biz360, Inc.

PHOENIX, Feb. 12 /PRNewswire/ -- Biz360, provider of external information analytics, today launched its first automated service, Market360, at IDG's DEMO 2001 Conference. With the unveiling of Market360, businesses can now bring meaningful measurement to such slippery concepts as buzz and mindshare. Biz360 today also announced Market360's first customers, networking giant, 3Com and e-Learning leader, DigitalThink.

Full Text:

Market360 Measures and Improves Public Relations Effectiveness;

First Customers Include 3Com and DigitalThink

PHOENIX, Feb. 12 /PRNewswire/ -- Biz360, provider of external information analytics, today launched its first automated service, Market360, at IDG's DEMO 2001 Conference. With the unveiling of Market360, businesses can now bring meaningful measurement to such slippery concepts as buzz and mindshare. Biz360 today also announced Market360's first customers, networking giant, 3Com and e-Learning leader, DigitalThink.

Producing metrics to measure information external to a company, such as media coverage, used to be a time-consuming manual process. In most cases, any analysis gleaned from reading piles of press clippings was old before complete and only pieces of an answer were ever found. Market360 changes this entirely.

"Public relations is one of the most important facets of a company's marketing effort," said Derek Gordon, Director of Corporate Communications of DigitalThink. "With the amazing analytics we get from Market360, we can now, for the first time, measure our PR effectiveness, see how we're doing compared with our competitors, and better target the writers and market influencers in our industry."

Market360 delivers analytics to objectively measure the effectiveness of PR campaigns, buzz, sentiment and mindshare. With intelligent metrics on the past and present, Market360 enhances targeting and goal setting for future PR activities. In addition, Market360 improves productivity in the day-to-day activities of PR and other marketing professionals. Since Market360 analyzes content in real time, marketing tactics can be adjusted and tuned in a more timely fashion with better results.

"Biz360 is a media roadmap. It tells 3Com where we are in the media and helps us chart where we're going next," said Brian Johnson, Director of Corporate Media Relations for 3Com Corporation. "Just like its name implies, it gives us a 360 degree view of our media profile. Biz360 lets us precisely assess the volume and quality of coverage we've earned."

Market360 automatically analyzes "raw" external data to deliver the reports PR and Marketing Professionals need most: Mindshare; prominence; ad-equivalency and message effectiveness. Market360 is customizable and reports are quickly modifiable based on many key criteria, e.g. time and competitor and key target information sources. Reports are also scalable, both in scope and depth. Users can start from the top line view and drill back to the original source article or posting as important trends are spotted.

"Biz360 shines a spotlight into the dark[illegible]s of marketing and public relations me[illegible]rement," said Chris Shipley, Executive Producer, DEMO Conferences. "Using the Market360 solution, marketing departments finally have the information they need to measure the effectiveness of public relations efforts and marketing campaigns. DEMO sets itself apart by launching companies with unique solutions to satisfy a proven need; therefore I'm very pleased to introduce Biz360 as a promising up-and-coming company at DEMO 2001."

About Biz360 (www.biz360.com)

Biz360 is a provider of automated analytic solutions for external company information. Biz360 solutions measure information previously thought unmeasurable like industry buzz, company mindshare and technology trends. Biz360 solutions combine the power of business analytics with the breadth of Internet content to deliver completely new, real-time market information and analysis services. Biz360's first solution, Market360, is directed to Marketing executives, Public Relations departments and PR agencies. Biz360 is a privately held company located in San Mateo, CA and is funded by Granite Ventures (formerly Hambrecht & Quist Venture Associates) and Adobe Ventures. SOURCE Biz360, Inc.

[Reference]
Message No: Industry: COMPUTER/ELECTRONICS; BANKING/FINANCIAL SERVICES;

Market360 Dashboard

Market Impact — Edit

[illegible]	$818,588
[illegible]	43,626,767
[illegible]	218,417
[illegible]	48

Mention Momentum

	Last 7 Days	Change
[illegible]	15	30%
[illegible]	13	34%
[illegible]	9	30%
[illegible]	7	-12%
[illegible]	4	0%
[illegible]	7	-61%

Market Impact — Edit

Top Authors — Edit

[illegible]	20
[illegible]	13
[illegible]	6
[illegible]	6
[illegible]	3

Top Publications — Edit

CNET [illegible]	17
[illegible]	15
Wall [illegible]	15
The [illegible]	15
Small [illegible] Magazine	13

Company Headlines — Email | Edit

Company

- U.S. Bancorp Piper Jaffray Announces Investment Opinion on Company; Initiates Coverage of Company with a Buy Rating and a $25 Price Target
  Feb 09 01 | News | The Street.com
- Diplomatic Planet announces Competitor E, Nobel Learning Systems and Company Among Consensus Picks for the eLearning 100
  Feb 07 01 | Press Release | Business Wire
- Thomas Weisel Partners Unveils Growth Stock Index
  Feb 01 01 | News | Yahoo! News
- Bain and Company Deliver Training Solutions
  Jan 31 01 | News | 2DNet
- AdvanceWork International, a Mobile-Learning Provider to Global Corporations Raises $3 Million in Funding
  Jan 29 01 | Press Release | Business Wire

Competitor's Headlines — Email | Edit

Competitor A

- Hot Products – October 1999
  Jan 23 01 | News | KMWorld.com
- Filling that ever-expanding reservoir of knowledge – A report on the KMWorld 2000 conference
  Jan 23 01 | News | KMWorld.com
- Competitor A sheds 47 percent of value
  Jan 17 01 | News | BizJournals.com
- Competitor A warns of deeper loss, cost-cutting moves
  Jan 17 01 | News | BizJournals.com
- Competitor A's shares plummet on dismal 4th-quarter news
  Jan 17 01 | News | Seattle Post-Intelligencer

Competitor B

- e-learning alliance

Search — Edit

Search

- (•) Keyword
- ( ) Author
- ( ) Publication

Watchwords — Email

1. Alliance
2. E-Learning

Missed Opportunities — Email

Count 24

My Reports

- Sentiment Over Time
- Message Effectiveness
- People vs. Press
- Mindshare Over Time – 2001
- Zero Mindshare Author

Source: Biz360, Inc.